On Convergence of Adam for Stochastic Optimization under Relaxed Assumptions

Yusu Hong
Center for Data Science
and School of Mathematical Sciences
Zhejiang University
yusuhong@zju.edu.cn

Junhong Lin[*]
Center for Data Science
Zhejiang University
junhong@zju.edu.cn

Abstract

In this paper, we study Adam in non-convex smooth scenarios with potential unbounded gradients and affine variance noise. We consider a general noise model which governs affine variance noise, bounded noise, and sub-Gaussian noise. We show that Adam with a specific hyper-parameter setup can find a stationary point with a $\mathcal{O}(\text{poly}(\log T)/\sqrt{T})$ rate in high probability under this general noise model where T denotes total number iterations, matching the lower rate of stochastic first-order algorithms up to logarithm factors. We also provide a probabilistic convergence result for Adam under a generalized smooth condition which allows unbounded smoothness parameters and has been illustrated empirically to capture the smooth property of many practical objective functions more accurately.

1 Introduction

Since its introduction by [33], the Stochastic Gradient Descent (SGD): $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t \boldsymbol{g}_t$ has achieved significant success in solving the unconstrained stochastic optimization problems:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}), \quad \text{where} \quad f(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\xi}}[f_{\boldsymbol{\xi}}(\boldsymbol{x},\boldsymbol{\xi})], \tag{1}$$

where $\boldsymbol{\xi}$ is a random variable, $\boldsymbol{g}_t$ is the stochastic gradients and η_t is the step-size. From then on, numerous literature focused on the convergence behavior of SGD in various scenarios. Several studies focused on the non-convex smooth scenario where the stochastic gradient $g(\boldsymbol{x})$ is unbiased with affine variance noise, i.e., for some constants $\sigma_0, \sigma_1 \geq 0$ and all $\boldsymbol{x} \in \mathbb{R}^d$,

$$\mathbb{E}[\|g(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^2] \leq \sigma_0^2 + \sigma_1^2 \|\nabla f(\boldsymbol{x})\|^2. \tag{2}$$

Under the noise assumption (2), [3] provided an almost-sure convergence bound for SGD. [4] proved that SGD could reach a stationary point with a $\mathcal{O}(1/\sqrt{T})$ rate when step-sizes are tuned by problem-parameters such as the smooth parameter L. The theoretical result also revealed that the analysis of SGD under (2) is not essentially different from the bounded noise case [17].

In the popular field of deep learning, a range of variants based on SGD, known as adaptive gradient methods have emerged. These methods employ the past gradients to adaptively tune their step-sizes and are preferred to SGD for minimizing various objective functions due to their efficiency. Among these methods, Adam [22] has been one of the most effective methods empirically. Generally speaking, Adam absorbs some key ideas from previous adaptive methods such as AdaGrad [12, 36] and RMSProp [37] while adding more unique structures. It combines the exponential moving average mechanism from RMSProp and meanwhile adds the heavy-ball style momentum [29] and two unique

[*] The corresponding author is Junhong Lin.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Table 1: Comparison for existing Adam analyses with ours.

	FCT	Grad.	Noise	Smooth	β_1, β_2	ϵ	Conv. Rate	Conv. Type
[48]	✗	Bounded	Bounded	L	$1-\beta_2 \le c\epsilon^2$	$\text{poly}(\frac{1}{\epsilon})$	$\frac{1}{T}+\sigma^2$	$\mathbb{E}$
[7]	✗	Bounded	Bounded	L	$\beta_{1,t} < \beta_1, \beta_2 = 1-\frac{1}{T}$	-	$\frac{1}{\sqrt{T}}$	$\mathbb{E}$
[57]	✗	Bounded	-	L	$\beta_2 = 1-\frac{c}{T}$	$\text{poly}(\log\frac{1}{\epsilon})$	$\frac{1}{\sqrt{T}}$	$\mathbb{E}$
[34]	✗	-	Finite Sum Affine	L	$T(\beta_1,\beta_2) \to 0^1$	-	-	$\mathbb{E}$
[10]	✗	Bounded	Bounded	L	$\beta_1 < \beta_2, \beta_2 = 1-\frac{1}{T}$	$\text{poly}(\log\frac{1}{\epsilon})$	$\frac{1}{\sqrt{T}}$	$\mathbb{E}$
[18]	✗	Bounded	**Affine**	L	$\beta_1 = 1-\frac{c}{\sqrt{T}}$	$\text{poly}(\frac{1}{\epsilon})$	$\frac{1}{\sqrt{T}}$	$\mathbb{E}$
[52]	✗	-	Finite Sum Affine	L	$\beta_1 < \sqrt{\beta_2}, \beta_2 = 1-\frac{c}{T}^{[3]}$	-	$\frac{1}{\sqrt{T}}$	$\mathbb{E}$
[41]	✗	-	Finite Sum Affine	$\boldsymbol{(L_0, L_1)}$	$\beta_1 < \sqrt{\beta_2}$	-	-	$\mathbb{E}$
[24]	✓	-	Sub-Gaussian	$\boldsymbol{(L_0, L_1)}$	$\beta_1 = 1-\frac{c}{\sqrt{T}}$	$\frac{1}{\sqrt{\epsilon}}$	$\frac{1}{\sqrt{T}}$	**w.h.p.**
[39]	✗	-	Coordinate-wise Affine	L	$\beta_1 = b\sqrt{\beta_2}, \beta_2 = 1-\frac{c}{T}$	$\text{poly}(\log\frac{1}{\epsilon})$	$\frac{1}{\sqrt{T}}$	$\mathbb{E}$
[19]	✓	-	Coordinate-wise Affine	L	$\beta_1 < \beta_2, \beta_2 = 1-\frac{1}{T}$	$\text{poly}(\log\frac{1}{\epsilon})$	$\frac{1}{\sqrt{T}}$	**w.h.p.**
Thm. 3.1	✓	-	**Affine**	L	$\beta_1 < \beta_2, \beta_2 = 1-\frac{c}{T}$	$\text{poly}(\log\frac{1}{\epsilon})$	$\frac{1}{\sqrt{T}}$	**w.h.p.**
Thm. 4.1	✓	-	**Affine**	$\boldsymbol{(L_0, L_1)}$	$\beta_1 < \beta_2, \beta_2 = 1-\frac{c}{T}$	$\text{poly}(\log\frac{1}{\epsilon})$	$\frac{1}{\sqrt{T}}$	**w.h.p.**

[1] [34] requires $T(\beta_1, \beta_2) = \mathcal{O}\left(\frac{\beta_1}{\beta_2^n}\left(\frac{1-\beta_1}{1-\beta_1^n}+1\right)\right) \to 0$, which seems could only achieve when $\beta_1 = 0$.

[2] Though not explicitly stated, the results in (Zhang et al., 2022) could imply convergence to the stationary point when with some calculations.

[3] "FCT" refers to "full corrective terms". The "Conv. rate" column presents the convergence rate omitting logarithm factors.

corrective terms. This unique structure leads to a huge success for Adam in practical applications but at the same time brings more challenges to the theoretical analysis.

Considering the significance of affine variance noise and Adam in both theoretical and empirical fields, it's natural to question whether Adam can find a stationary point at a rate comparable to SGD under the same smooth condition and (2). Earlier researches [14, 40, 2] have shown that AdaGrad-Norm, a scalar version of AdaGrad, can find a stationary point at the same rate as SGD, not tuning step-sizes based on problem-parameters. Moreover, they addressed an essential challenge brought by the correlation of adaptive step-sizes and noise from (2) which does not appear in SGD's cases. However, since AdaGrad-Norm applies a cumulative step-sizes mechanism which is rather different from the exponential moving average step-sizes in Adam, the analysis for AdaGrad-Norm could not be trivially extended to Adam. Furthermore, the coordinate-wise step-size architecture of Adam, rather than the unified step-size for all coordinates in AdaGrad-Norm, brings more challenge when considering (2). In affine variance noise landscape, existing literature could only ensure the Adam's convergence with random-reshuffling scheme under certain parameter restrictions [52, 41], or deduce the convergence at the expense of requiring bounded gradient assumption and using problem-parameters to tune the step-sizes [18], both of which ignored the corrective terms. Some other works proved convergence to a stationary point by altering the original Adam algorithm such as removing certain corrective terms and modifying (2) to a stronger coordinate-wise variant [39, 19].

To the best of our knowledge, existing research has not yet fully confirmed the convergence of Adam under affine variance noise. To address this gap, we conduct an in-depth analysis and prove that Adam with the right parameter can find a stationary point in high probability. We assume a milder noise model (detailed in Assumption (A3)), covering almost surely affine variance noise, the bounded noise, and sub-Gaussian noise. We show that the convergence rate can reach at $\mathcal{O}\left(\text{poly}(\log T)/\sqrt{T}\right)$ matching the lower rate in [1] up to logarithm factors. Our proof employs the descent lemma over the introduced proxy iterative sequence and adopts techniques related to the new proxy step-sizes and error decomposition. Based on this, we are able to handle the correlation between stochastic gradients and adaptive step-sizes and transform the first-order term from the descent lemma into the gradient norm.

Finally, we apply the analysis to the (L_0, L_q)-smooth condition [51]. Several researchers have found empirical evidence of objective functions satisfying (L_0, L_q)-smoothness but out of L-smoothness range, especially in large-scale language models [49, 38, 11, 8]. Theoretical analysis of adaptive methods under this relaxed condition is more complicated and needs further nontrivial proof techniques. Also, prior knowledge of problem-parameters to tune step-sizes is needed, as indicated by the counter-examples from [40] for the AdaGrad. Existing works [13, 40] obtained a convergence bound for AdaGrad-Norm with (2), and [24] considered Adam with sub-Gaussian noise. In this paper, we provide a probabilistic convergence result for Adam with the affine variance noise and the generalized smoothness condition.

Algorithm 1 Adam

```
**Input:** Horizon $T$, $\boldsymbol{x}_1 \in \mathbb{R}^d$, $\beta_1, \beta_2 \in [0,1)$, $\boldsymbol{m}_0 = \boldsymbol{v}_0 = \boldsymbol{0}_d$, $\eta, \epsilon > 0$, $\boldsymbol{\epsilon} = \epsilon \boldsymbol{1}_d$
**for** $s = 1, \cdots, T$ **do**
    Draw a new sample $\boldsymbol{z}_s$ and generate $\boldsymbol{g}_s = g(\boldsymbol{x}_s, \boldsymbol{z}_s)$;
    $\boldsymbol{m}_s = \beta_1 \boldsymbol{m}_{s-1} + (1-\beta_1)\boldsymbol{g}_s$;
    $\boldsymbol{v}_s = \beta_2 \boldsymbol{v}_{s-1} + (1-\beta_2)\boldsymbol{g}_s^2$;
    $\eta_s = \eta\sqrt{1-\beta_2^s}/(1-\beta_1^s)$, $\boldsymbol{\epsilon}_s = \boldsymbol{\epsilon}\sqrt{1-\beta_2^s}$;
    $\boldsymbol{x}_{s+1} = \boldsymbol{x}_s - \eta_s \cdot \boldsymbol{m}_s / \left(\sqrt{\boldsymbol{v}_s} + \boldsymbol{\epsilon}_s\right)$;
**end for**
```

We also refer readers to see the main contributions of our works and comparisons with the existing works in Table 1.

Notations We use $[T]$ to denote the set $\{1, 2, \cdots, T\}$ for any positive integer T, $\|\cdot\|$, $\|\cdot\|_1$ and $\|\cdot\|_\infty$ to denote l_2-norm, l_1-norm and l_∞-norm respectively. $a \sim \mathcal{O}(b)$ and $a \le \mathcal{O}(b)$ denote $a = C_1 b$ and $a \le C_2 b$ for some positive universal constants C_1, C_2, and $a \le \tilde{\mathcal{O}}(b)$ denotes $a \le \mathcal{O}(b)\text{poly}(\log b)$. $a \lesssim b$ denotes $a \le \mathcal{O}(b)$. For any vector $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{x}^2$ and $\sqrt{\boldsymbol{x}}$ denote coordinate-wise square and square root respectively. $\boldsymbol{x}_i$ denotes the i-th coordinate of $\boldsymbol{x}$. For any two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we use $\boldsymbol{x} \odot \boldsymbol{y}$ and $\boldsymbol{x}/\boldsymbol{y}$ to denote the coordinate-wise product and quotient respectively. $\boldsymbol{0}_d$ and $\boldsymbol{1}_d$ represent zero and one d-dimensional vectors respectively.

2 Problem set up and algorithm

We consider unconstrained stochastic optimization (1) over $\mathbb{R}^d$ with l_2-norm. The objective function $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable. Given $\boldsymbol{x} \in \mathbb{R}^d$, we assume a gradient oracle that returns a random vector $g(\boldsymbol{x}, \boldsymbol{z}) \in \mathbb{R}^d$ dependent by the random sample $\boldsymbol{z}$. The true gradient of f at $\boldsymbol{x}$ is denoted by $\nabla f(\boldsymbol{x}) \in \mathbb{R}^d$.

Assumptions We make the following assumptions throughout the paper.

- (**A1**) Bounded below: There exists $f^* > -\infty$ such that $f(\boldsymbol{x}) \ge f^*, \forall \boldsymbol{x} \in \mathbb{R}^d$;
- (**A2**) Unbiased estimator: The gradient oracle provides an unbiased estimator of $\nabla f(\boldsymbol{x})$, i.e., $\mathbb{E}_{\boldsymbol{z}}\left[g(\boldsymbol{x}, \boldsymbol{z})\right] = \nabla f(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathbb{R}^d$;
- (**A3**) Generalized affine variance noise: The gradient oracle satisfies that there are some constants $\sigma_0, \sigma_1 > 0, p \in [0, 4)$, $\mathbb{E}_{\boldsymbol{z}}\left[\exp\left(\frac{\|g(\boldsymbol{x},\boldsymbol{z}) - \nabla f(\boldsymbol{x})\|^2}{\sigma_0^2 + \sigma_1^2\|\nabla f(\boldsymbol{x})\|^p}\right)\right] \le \exp(1), \forall \boldsymbol{x} \in \mathbb{R}^d$.

The first two assumptions are standard in the stochastic optimization. The third assumption provides a mild noise model that covers the almost surely bounded noise and sub-Gaussian noise. Moreover, it's more general than almost surely affine variance noise as follows

$$\|g(\boldsymbol{x}, \boldsymbol{z}) - \nabla f(\boldsymbol{x})\|^2 \le \sigma_0^2 + \sigma_1^2\|\nabla f(\boldsymbol{x})\|^2, a.s., \tag{3}$$

and enlarge the range of p to $[0, 4)$. Assumption (**A3**) with $p = 2$ and (3) are also utilized in [2] to establish high probability results for AdaGrad-Norm. It represents a stronger condition than the expected version of (2) that is commonly employed for deriving the expected convergence of algorithms. However, almost surely assumption enables the derivation of stronger high-probability convergence guarantees for algorithms, while still ensuring expected convergence.

The affine noise variance assumption is important for machine learning applications with feature noise (including missing features) [15, 21], in robust linear regression [45], and generally whenever the model parameters are multiplicatively perturbed by noise (e.g., a multilayer network, where noise from a previous layer multiplies the parameters in subsequent layers). We refer interested readers to see e.g., [3, 45, 4, 14, 40, 2] for more discussions about the affine variance noise.

Adam For the stochastic optimization problem, we study Algorithm 1, which is an equivalent form of Adam [22] with the two corrective terms for $\boldsymbol{m}_s$ and $\boldsymbol{v}_s$ included into η_s for notation simplicity.

The iterative relationship in Algorithm 1 can be also written as for any $s \in [T]$,

$$\boldsymbol{x}_{s+1} = \boldsymbol{x}_s - \eta_s(1-\beta_1) \cdot \frac{\boldsymbol{g}_s}{\sqrt{\boldsymbol{v}_s} + \boldsymbol{\epsilon}_s} + \beta_1 \cdot \frac{\eta_s(\sqrt{\boldsymbol{v}_{s-1}} + \boldsymbol{\epsilon}_{s-1})}{\eta_{s-1}(\sqrt{\boldsymbol{v}_s} + \boldsymbol{\epsilon}_s)} \odot (\boldsymbol{x}_s - \boldsymbol{x}_{s-1}), \tag{4}$$

where we let $\boldsymbol{x}_0 = \boldsymbol{x}_1$ and $\eta_0 = \eta$. (4) plays a key role in the convergence analysis, showing that Adam incorporates a heavy-ball style momentum and dynamically adjusts its momentum through β_1 and β_2, along with adaptive step-sizes. This inspires us to learn from some classical analysis methods for algorithms with momentum and provides some new estimations to fit in with the adaptive property.

3 Convergence of Adam with smooth objective functions

In this section, we assume that the objective function f is L-smooth satisfying that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\| \le L\|\boldsymbol{y} - \boldsymbol{x}\|. \tag{5}$$

We then show that Adam has the following high probability results.

Theorem 3.1. *Let $T \ge 1$ and $\{\boldsymbol{x}_s\}_{s\in[T]}$ be the sequence generated by Algorithm 1. If Assumptions (A1)-(A3) hold, and the hyper-parameters satisfy that*

$$0 \le \beta_1 < \beta_2 < 1, \quad \beta_2 = 1 - c/T, \quad \eta = C_0\sqrt{1-\beta_2}, \quad \epsilon = \epsilon_0\sqrt{1-\beta_2}, \tag{6}$$

for some constants $c, C_0 > 0$ and $\epsilon_0 > 0$, then for any given $\delta \in (0, 1/2)$, it holds that with probability at least $1 - 2\delta$,

$$\frac{1}{T}\sum_{s=1}^{T} \|\nabla f(\boldsymbol{x}_s)\|^2 \le \mathcal{O}\left\{ G^2 \left(\sqrt{\frac{\sigma_0^2 + \sigma_1^2 G^p + G^2}{T}} + \frac{\epsilon_0}{T} \right) \log\left(\frac{T}{\delta}\right) \right\},$$

where G^2 is defined by the following order with respect to T, ϵ_0, δ:[2]

$$G^2 \sim \mathcal{O}\left(\log^{\frac{3}{2}\max\{2, \frac{4}{4-p}\}} \left(\frac{T}{\epsilon_0\delta}\right) \right). \tag{7}$$

Theorem 3.1 provides the nearly optimal convergence rate $\mathcal{O}\left(\text{poly}(\log T)/\sqrt{T}\right)$ to find a stationary point when setting the parameter probably: $\beta_2 = 1 - \mathcal{O}(1/T)$. It's worth noting that the setting requires β_2 to be closed enough to 1 when T is sufficiently large, which roughly aligns with the typical setting in [22, 57, 10, 39].

For a more detailed comparison of our results to existing works, including assumptions, convergence rate, and dependency, we refer readers to Table 1.

4 Convergence of Adam with generalized smooth objective functions

In this section, we study the convergence behavior of Adam in the generalized smooth case. We first provide some necessary introduction to the generalized smooth condition.

4.1 Generalized smoothness

For a differentiable objective function $f : \mathbb{R}^d \to \mathbb{R}$, we consider the following (L_0, L_q)-smoothness condition: there exist constants $q \in [0, 2)$ and $L_0, L_q > 0$, satisfying that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ with $\|\boldsymbol{x} - \boldsymbol{y}\| \le 1/L_q$,

$$\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\| \le (L_0 + L_q\|\nabla f(\boldsymbol{x})\|^q)\, \|\boldsymbol{x} - \boldsymbol{y}\|. \tag{8}$$

The generalized smooth condition was originally put forward by [51] for any twice differentiable function f satisfying that

$$\|\nabla^2 f(\boldsymbol{x})\| \le L_0 + L_1\|\nabla f(\boldsymbol{x})\|. \tag{9}$$

[2] Note that we hide the constants β_1, β_2 inside $\mathcal{O}$. The detailed expression of G^2 could be found in (52) from Appendix.

It has been proved that a lot of objective functions in experimental areas satisfy (9) but out of L-smoothness range, especially in training large language models, see e.g., Figure 1 in [51] and [8].

To better understand the theoretical significance of the generalized smoothness, [49] provided an alternative form in (8) with $q = 1$, only requiring f to be differentiable. They showed that (8) is sufficient to elucidate the convergence of gradient-clipping algorithms.

There are three key reasons for opting for (8). Firstly, considering our access is limited to first-order stochastic gradients, it's logical to only assume that f is differentiable. Second, as pointed out by Lemma A.2 in [49] and Proposition 1 in [13], (8) and (9) are equivalent up to constant factors when f is twice differentiable considering $q = 1$. Thus, (8) covers a broader range of functions than (9). Finally, it's easy to verify that (8) is strictly weaker than L-smoothness. A concrete example is that the simple function $f(x) = x^4, x \in \mathbb{R}$ does not satisfy any global L-smoothness but (8). Moreover, the expanded range of q to $[0, 2)$ is necessary as all univariate rational functions $P(x)/Q(x)$, where P, Q are polynomials and double exponential functions $a^{(b^x)}$ with $a, b > 1$ are (L_0, L_q)-smooth with $1 < q < 2$ (see [24, Proposition 3.4]). We refer interested readers to see [51, 49, 13, 24] for more discussions of concrete examples of generalized smoothness.

4.2 Convergence result

We then provide the high probability convergence result of Adam with (L_0, L_q)-smoothness condition as follows.

Theorem 4.1. *Let $T \geq 1$ and $\delta \in (0, 1/2)$. Suppose that $\{\boldsymbol{x}_s\}_{s\in[T]}$ is a sequence generated by Algorithm 1, f is (L_0, L_q)-smooth satisfying* (8)*, Assumptions (A1)-(A3) hold, and the parameters satisfy*

$$\begin{aligned}
&0 \leq \beta_1 < \beta_2 < 1, \quad \beta_2 = 1 - c/T, \quad \epsilon = \epsilon_0\sqrt{1-\beta_2}, \quad \eta = \tilde{C}_0\sqrt{1-\beta_2}, \\
&\tilde{C}_0 \leq \min\left\{E_0, \frac{E_0}{\mathcal{H}}, \frac{E_0}{\mathcal{L}}, \sqrt{\frac{\beta_2(1-\beta_1)^2(1-\beta_1/\beta_2)}{4L_q^2 d}}\right\},
\end{aligned} \tag{10}$$

where $c, \epsilon_0, E_0, \tilde{C}_0 > 0$ are constants, $\hat{H}$ is controlled by $\mathcal{O}\left(\log\left(\frac{T}{\epsilon_0\delta}\right)\right)$ [3], and $H, \mathcal{H}, \mathcal{L}$ are defined as

$$\begin{aligned}
&H := L_0/L_q + \left(4L_q\hat{H}\right)^q + \left(4L_q\hat{H}\right)^{\frac{q}{2-q}} + \left(4L_0\hat{H}\right)^{\frac{q}{2}} + 4L_q\hat{H} + \left(4L_q\hat{H}\right)^{\frac{1}{2-q}} + \sqrt{4L_0\hat{H}}, \\
&\mathcal{H} := \sqrt{2(\sigma_0^2 + \sigma_1^2 H^p + H^2)\log\left(\frac{\mathrm{e}T}{\delta}\right)}, \quad \mathcal{L} := L_0 + L_q\left(H^q + H + \frac{L_0}{L_q}\right)^q.
\end{aligned} \tag{11}$$

Then it holds that with probability at least $1 - 2\delta$,

$$\frac{1}{T}\sum_{s=1}^{T}\|\nabla f(\boldsymbol{x}_s)\|^2 \leq \mathcal{O}\left\{\frac{\hat{H}}{\tilde{C}_0}\left(\sqrt{\frac{\sigma_0^2 + \sigma_1^2 H^p + H^2}{T}} + \frac{\epsilon_0}{T}\right)\log\left(\frac{T}{\delta}\right)\right\}. \tag{12}$$

Note that in the above theorem, the order of $\log T$ in $\hat{H}$ and the final convergence bound is better than the one in Theorem 3.1 under the same noise assumption. This better dependency comes from the expense of using problem parameters to tune step-size $\tilde{C}_0$. Since $\hat{H}$ is logarithm order of T, $H, \mathcal{H}, \mathcal{L}$ are both polynomial logarithm order of T and the final convergence rate in (12) is also polynomial logarithm order of T. Note that $\tilde{C}_0 \leq \mathcal{O}(1/\text{poly}(\log T))$ from (10) when $T \gg d$. Hence, when T is large enough, a possible optimal setting is that $\eta = c_1/(\sqrt{T}\text{poly}(\log T))$ for some constant $c_1 > 0$, which roughly matches the typical setting as mentioned before.

5 Related works

There is a large amount of works on stochastic approximations (or online learning algorithms) and adaptive variants, e.g., [5, 35, 47, 28, 12, 4, 6, 26, 54] and the references therein. In this section, we will discuss the most related works and make a comparison with our main results.

[3] The specific definition of $\hat{H}$ can be found in (82) from Appendix.

5.1 Convergence with affine variance noise and its variants

We mainly list previous literature considering (2) over non-convex smooth scenario. [3] provided an asymptotic convergence result for SGD with (2). In terms of non-asymptotic results, [4] proved the convergence of SGD, illustrating that the analysis was non-essentially different from the bounded noise case from [17].

In the adaptive methods field, [14] studied convergence of AdaGrad-Norm with (2), pointing out that the analysis is more challenging than the bounded noise and bounded gradient case in [43]. They provided a convergence rate of $\tilde{\mathcal{O}}(1/\sqrt{T})$ without knowledge of problem parameters, and further improved the bound adapting to the noise level: when $\sigma_1 \sim \mathcal{O}(1/\sqrt{T})$,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{x}_t)\|^2 \le \tilde{\mathcal{O}}\left(\frac{\sigma_0}{\sqrt{T}}+\frac{1}{T}\right). \tag{13}$$

(13) matches exactly with SGD's case [4], showing a fast rate of $\tilde{\mathcal{O}}(1/T)$ when σ_0 is sufficiently low. Later, [40] proposed a deep analysis framework obtaining (13) with a tighter dependency to T and not requiring any restriction over σ_1. They further obtained the same rate for AdaGrad under a stronger coordinate-wise version of (2): for all $i \in [d]$,

$$\mathbb{E}_{\boldsymbol{z}}|\boldsymbol{g}(\boldsymbol{x},\boldsymbol{z})_i - \nabla f(\boldsymbol{x})_i|^2 \le \sigma_0^2 + \sigma_1^2|\nabla f(\boldsymbol{x})_i|^2. \tag{14}$$

[2] obtained a probabilistic convergence rate for AdaGrad-Norm with (3) using a novel induction argument to estimate the function value gap without any requirement over σ_1 as well.

In the analysis of Adam, a line of works [34, 52, 41] considered Adam without corrective terms for finite-sum objective functions under different regimes while possibly incorporating natural random shuffling technique. They could ensure that this variant converged to a bounded region where

$$\min_{t\in[T]}\mathbb{E}\left[\min\{\|\nabla f(\boldsymbol{x}_t)\|, \|\nabla f(\boldsymbol{x}_t)\|^2\}\right] \lesssim \frac{\log T}{\sqrt{T}} + C_1\sigma_0 \tag{15}$$

under the affine growth condition which is equivalent to (2). Though not explicitly concluded, when setting $\beta_2 = 1-\mathcal{O}(1/T)$, [52]'s work can also ensure a convergence rate of order $\tilde{\mathcal{O}}(1/\sqrt{T})$ under certain settings. Besides, both [20] and [18] provided convergence bounds allowing for large heavy-ball momentum parameter that aligns more closely with practical settings. However, they relied on the assumption for step-sizes where $C_l \le \|\frac{1}{\sqrt{\boldsymbol{v}_t}+\boldsymbol{\epsilon}_t}\|_\infty \le C_u, \forall t \in [T]$. [39] and [19] used distinct methods to derive convergence bounds in expectation and high probability respectively, without relying on bounded gradients. Both studies achieved a convergence rate of the form in (13) for Adam ignoring the corrective terms. [19] further achieved a $\tilde{\mathcal{O}}(1/\sqrt{T})$ rate for Adam. However, the two works only studied coordinate-wise affine variance noise.

In this paper, we derive a stronger high probability convergence rate for Adam with original corrective terms, relying on an almost surely noise assumption. The noise model is general enough to cover bounded noise, sub-Gaussian noise, and (coordinate-wise) affine variance noise. Although we consider a stronger almost surely assumption, our probabilistic convergence result is also stronger than the expected convergence.

5.2 Convergence with generalized smoothness

The generalized smooth condition was first proposed for twice differentiable functions by [51] (see (9)) to explain the acceleration mechanism of gradient-clipping. This assumption was extensively confirmed in experiments of large-scale language models [51]. Later, [49] further relaxed it to a more general form in (8) allowing for first-order differentiable functions. Subsequently, a series of works [30, 53, 32] studied different algorithms' convergence under this condition.

In the field of adaptive methods, [13] provided a convergence bound for AdaGrad-Norm assuming (2) and (8) with $q = 1$, albeit requiring $\sigma_1 < 1$. Based on the same conditions, [40] improved the convergence rate to the form in (13) without restriction on σ_1. [41] explored how Adam without corrective terms behaves under generalized smoothness with $q = 1$ and (2). However, they could only assert convergence to a bounded region as shown in (15). [8] showed that an Adam-type algorithm converges to a stationary point under a stronger coordinate-wise generalized smooth condition.

Recently, [24] provided a novel framework to derive high probability convergence bound for Adam under the generalized smooth and sub-Gaussian noise case.

In this paper, we consider a more general noise setup and investigate Adam's convergence under the generalized smooth landscape. We prove that Adam is powerful enough to find a stationary point with properly tuned step-sizes even under these relaxed assumptions. Moreover, the convergence rate is not harmed by the relaxation of noise and smoothness, matching the optimal $\mathcal{O}(1/\sqrt{T})$ rate up to logarithm factors.

5.3 Convergence of Adam

Adam was first proposed by [22] with empirical studies and theoretical results on online convex learning. The original proof of convergence in [22] was later shown by [31] to contain gaps. [31] and the subsequent work [42] also showed that for a range of momentum parameters chosen independently with the problem instance, Adam does not necessarily converge even for convex objectives. Many works have focused on its convergence behavior in non-convex smooth fields. A series of works studied Adam ignoring corrective terms, all requiring a uniform bound for gradients' norm. Among these works, [48] demonstrated that Adam can converge within a specific region if step-sizes and decay parameters are determined properly by the smooth parameter. [9] proposed a convergence result to a stationary point and required all stochastic gradients must keep the same sign. To circumvent this requirement, [57] introduced a convergence bound only requiring hyper-parameters to satisfy specific conditions. [10] conducted a simple proof and further improved the dependency on the heavy-ball momentum parameter. Recently, [55] introduced Nesterov-like acceleration into Adam and AdamW [27] indicating their superiority in convergence over the non-accelerated versions. For Adam-related works under (2) or generalized smoothness, we refer readers to Sections 5.1 and 5.2.

We also want to highlight that a series of works [23, 44, 50] investigated the geometry of Adam from an l_∞-norm perspective. [23] and [44] studied the geometry of Adam by regarding it as a variant of SignSGD and [50] showed that full-batch Adam converges towards a linear classifier that achieves the maximum l_∞-margin when the training data are linearly separable.

6 Proof sketch under the smooth case

In this section, we provide a proof sketch of Theorem 3.1 with some insights and proof novelty. Our proof borrows some ideas from [43, 10, 14, 2, 39, 19]. The detailed proof can be found in Appendix B.

Preliminary To start with, we let the stochastic gradient $\boldsymbol{g}_s = (g_{s,i})_i$, the true gradient $\nabla f(\boldsymbol{x}_s) = \bar{\boldsymbol{g}}_s = (\bar{g}_{s,i})_i$ and $\boldsymbol{\xi}_s = (\xi_{s,i})_i = \boldsymbol{g}_s - \bar{\boldsymbol{g}}_s$. We also let $\epsilon_s = \epsilon\sqrt{1-\beta_2^s}$ and thus $\boldsymbol{\epsilon}_s = \epsilon_s \mathbf{1}_d$. For any positive integer T and $\delta \in (0,1)$, we define $\mathcal{M}_T = \sqrt{\log(\mathrm{e}T/\delta)}$. We denote the adaptive part of the step-size as

$$\boldsymbol{b}_s := \sqrt{\boldsymbol{v}_s} + \boldsymbol{\epsilon}_s = \sqrt{\beta_2 \boldsymbol{v}_{s-1} + (1-\beta_2)\boldsymbol{g}_s^2} + \boldsymbol{\epsilon}_s. \tag{16}$$

We define two auxiliary sequences $\{\boldsymbol{p}_s\}_{s\ge 1}$ and $\{\boldsymbol{y}_s\}_{s\ge 1}$,

$$\boldsymbol{p}_1 = \mathbf{0}_d, \quad \boldsymbol{y}_1 = \boldsymbol{x}_1, \quad \boldsymbol{p}_s = \frac{\beta_1}{1-\beta_1}(\boldsymbol{x}_s - \boldsymbol{x}_{s-1}), \boldsymbol{y}_s = \boldsymbol{p}_s + \boldsymbol{x}_s, \forall s \ge 2. \tag{17}$$

We follow from [16, 46] which was used to prove the convergence of SGD with momentum and later applied to handle many variants of momentum-based algorithms. Recalling the iteration of $\boldsymbol{x}_s$ in (4), we reveal that $\boldsymbol{y}_s$ satisfies

$$\boldsymbol{y}_{s+1} = \boldsymbol{y}_s - \eta_s \cdot \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s} + \frac{\beta_1}{1-\beta_1}\left(\frac{\eta_s \boldsymbol{b}_{s-1}}{\eta_{s-1}\boldsymbol{b}_s} - \mathbf{1}_d\right) \odot (\boldsymbol{x}_s - \boldsymbol{x}_{s-1}). \tag{18}$$

In addition, given $T \ge 1$, we define, $\forall s \in [T]$,

$$G_s = \max_{j\in[s]} \|\bar{\boldsymbol{g}}_j\|, \mathcal{G}_T(s) = \mathcal{M}_T\sqrt{2\sigma_0^2 + 2\sigma_1^2 G_s^p + 2G_s^2}, \mathcal{G}_T = \mathcal{M}_T\sqrt{2\sigma_0^2 + 2\sigma_1^2 G^p + 2G^2}, \tag{19}$$

where G is as in Theorem 3.1. Both G_s and $\mathcal{G}_T(s)$ will serve as upper bounds for gradients' norm before time s. We will verify their importance in the later argument.

Starting from the descent lemma We fix the horizon T and start from the standard descent lemma of L-smoothness. Then, for any given $t \in [T]$, combining with (18) and summing over $s \in [t]$,

$$f(\boldsymbol{y}_{t+1}) \le f(\boldsymbol{x}_1) + \underbrace{\sum_{s=1}^{t} -\eta_s \left\langle \nabla f(\boldsymbol{y}_s), \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s} \right\rangle}_{\mathbf{A}} + \underbrace{\frac{\beta_1}{1-\beta_1} \sum_{s=1}^{t} \langle \Delta_s \odot (\boldsymbol{x}_s - \boldsymbol{x}_{s-1}), \nabla f(\boldsymbol{y}_s) \rangle}_{\mathbf{B}}$$

$$+ \underbrace{\frac{L}{2} \sum_{s=1}^{t} \left\| \eta_s \cdot \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s} - \frac{\beta_1}{1-\beta_1} (\Delta_s \odot (\boldsymbol{x}_s - \boldsymbol{x}_{s-1})) \right\|^2}_{\mathbf{C}}, \tag{20}$$

where we let $\Delta_s = \frac{\eta_s \boldsymbol{b}_{s-1}}{\eta_{s-1} \boldsymbol{b}_s} - \mathbf{1}_d$ and use $\boldsymbol{y}_1 = \boldsymbol{x}_1$ from (17). In what follows, we will estimate **A**, **B**, and **C** respectively.

Probabilistic estimations To proceed with the analysis, we next introduce two probabilistic estimations showing that the norm of the noises and a related summation of martingale difference sequence could be well controlled with high probability. We show that with probability at least $1 - 2\delta$, the following two inequalities hold simultaneously for all $t \in [T]$:

$$\|\boldsymbol{\xi}_t\|^2 \le \mathcal{M}_T^2 \left(\sigma_0^2 + \sigma_1^2 \|\bar{\boldsymbol{g}}_t\|^p \right), \quad \text{and} \tag{21}$$

$$-\sum_{s=1}^{t} \eta_s \left\langle \bar{\boldsymbol{g}}_s, \frac{\boldsymbol{\xi}_s}{\boldsymbol{a}_s} \right\rangle \le \frac{\mathcal{G}_T(t)}{4\mathcal{G}_T} \sum_{s=1}^{t} \eta_s \left\| \frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}} \right\|^2 + D_1 \mathcal{G}_T, \tag{22}$$

where D_1 is a constant defined in Lemma B.7 and $\boldsymbol{a}_s$ will be introduced later. In what follows, we always assume that (21) and (22) hold for all $t \in [T]$ and carry out our subsequent analysis with some deterministic estimations.

Estimating A We first decompose **A** as

$$\mathbf{A} = \underbrace{\sum_{s=1}^{t} -\eta_s \left\langle \bar{\boldsymbol{g}}_s, \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s} \right\rangle}_{\mathbf{A.1}} + \underbrace{\sum_{s=1}^{t} \eta_s \left\langle \bar{\boldsymbol{g}}_s - \nabla f(\boldsymbol{y}_s), \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s} \right\rangle}_{\mathbf{A.2}}.$$

Due to the correlation of the stochastic gradient $\boldsymbol{g}_s$ and the step-size $\eta_s / \boldsymbol{b}_s$, the estimating of **A.1** is challenging, as also noted in the analysis for other adaptive gradient methods, e.g., [43, 10, 14, 2, 39, 19]. To break this correlation, the so-called proxy step-size technique is introduced and variants of proxy step-size have been introduced in the related literature. However, to our best knowledge, none of these proxy step-sizes could be used in our analysis for Adam considering potential unbounded gradients under the noise model in Assumption (A3). In this paper, we construct a proxy step-size $\eta_s / \boldsymbol{a}_s$, with $\boldsymbol{a}_s$ relying on $\mathcal{G}_T(s)$ in (19), defined as for any $s \in [T]$,

$$\boldsymbol{a}_s = \sqrt{\beta_2 \boldsymbol{v}_{s-1} + (1-\beta_2) \left(\mathcal{G}_T(s) \mathbf{1}_d \right)^2} + \boldsymbol{\epsilon}_s. \tag{23}$$

With the so-called proxy step-size technique over $\eta_s / \boldsymbol{a}_s$ and $\boldsymbol{\xi}_s = \boldsymbol{g}_s - \bar{\boldsymbol{g}}_s$, we decompose **A.1** as

$$\mathbf{A.1} = -\sum_{s=1}^{t} \eta_s \left\| \frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}} \right\|^2 \underbrace{- \sum_{s=1}^{t} \eta_s \left\langle \bar{\boldsymbol{g}}_s, \frac{\boldsymbol{\xi}_s}{\boldsymbol{a}_s} \right\rangle}_{\mathbf{A}.1.1} + \underbrace{\sum_{s=1}^{t} \eta_s \left\langle \bar{\boldsymbol{g}}_s, \left(\frac{1}{\boldsymbol{a}_s} - \frac{1}{\boldsymbol{b}_s} \right) \boldsymbol{g}_s \right\rangle}_{\mathbf{A.1.2}}.$$

In the above decomposition, the first term serves as a descent term. **A.1.1** is now a summation of a martingale difference sequence which could be estimated by (22). **A.1.2** is regarded as an error term when introducing $\boldsymbol{a}_s$. However, due to the delicate construction of $\boldsymbol{a}_s$, the definition of local gradients' bound $\mathcal{G}_T(t)$, and using some basic inequalities, we show that

$$\mathbf{A.1.2} \le \frac{1}{4} \sum_{s=1}^{t} \eta_s \left\| \frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}} \right\|^2 + \frac{\eta \mathcal{G}_T(t) \sqrt{1-\beta_2}}{1-\beta_1} \sum_{s=1}^{t} \left\| \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s} \right\|^2.$$

The first RHS term can be eliminated with the descent term while the summation of the last term can be bounded by

$$\sum_{s=1}^{t}\left\|\frac{\boldsymbol{g}_s}{\boldsymbol{b}_s}\right\|^2 \vee \sum_{s=1}^{t}\left\|\frac{\boldsymbol{m}_s}{\boldsymbol{b}_s}\right\|^2 \vee \sum_{s=1}^{t}\left\|\frac{\boldsymbol{m}_s}{\boldsymbol{b}_{s+1}}\right\|^2 \vee \sum_{s=1}^{t}\left\|\frac{\hat{\boldsymbol{m}}_s}{\boldsymbol{b}_s}\right\| \lesssim \frac{d}{1-\beta_2}\log\left(\frac{T}{\beta_2^T}\right), \tag{24}$$

due to the step-size's adaptivity, the iterative relationship of the algorithm, the smoothness of the objective function, as well as (21). Here, $\hat{\boldsymbol{m}}_s = \frac{\boldsymbol{m}_s}{1-\beta_1^s}$.

Estimating B and C The key to estimate **B** is to decompose **B** as

$$\mathbf{B} = \underbrace{\frac{\beta_1}{1-\beta_1}\sum_{s=1}^{t}\langle \Delta_s \odot (\boldsymbol{x}_s - \boldsymbol{x}_{s-1}), \bar{\boldsymbol{g}}_s\rangle}_{\mathbf{B.1}} + \underbrace{\frac{\beta_1}{1-\beta_1}\sum_{s=1}^{t}\langle \Delta_s \odot (\boldsymbol{x}_s - \boldsymbol{x}_{s-1}), \nabla f(\boldsymbol{y}_s) - \bar{\boldsymbol{g}}_s\rangle}_{\mathbf{B.2}}.$$

To estimate **B.1**, we use the updated rule and further write $\Delta_s \odot (\boldsymbol{x}_s - \boldsymbol{x}_{s-1})$ as

$$\left(\frac{\eta_s}{\boldsymbol{b}_s} - \frac{\eta_s}{\boldsymbol{a}_s}\right) \odot \boldsymbol{m}_{s-1} + \left(\frac{\eta_s}{\boldsymbol{a}_s} - \frac{\eta_s}{\boldsymbol{b}_{s-1}}\right) \odot \boldsymbol{m}_{s-1} + (\eta_s - \eta_{s-1})\frac{\boldsymbol{m}_{s-1}}{\boldsymbol{b}_{s-1}},$$

and upper bound the three related inner products. Using some basic inequalities, the smoothness, (24), and some delicate computations, one can estimate the three related inner products, **B.2** and **C**, and thus get that

$$\mathbf{B} + \mathbf{C} \le \frac{1}{4}\sum_{s=1}^{t}\eta_s\left\|\frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}}\right\|^2 + (b_1\mathcal{G}_T(t) + b_2)\log\left(\frac{T}{\beta_2^T}\right),$$

where b_1 and b_2 are positive constants determined by $\beta_1, \beta_2, d, L, \eta$.

Bounding gradients through induction The last challenge comes from the potential unbounded gradients' norm. Plugging the above estimations into (20), we obtain that

$$f(\boldsymbol{y}_{t+1}) \le f(\boldsymbol{x}_1) + \left(\frac{\mathcal{G}_T(t)}{4\mathcal{G}_T} - \frac{1}{2}\right)\sum_{s=1}^{t}\eta_s\left\|\frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}}\right\|^2 + c_1\mathcal{G}_T + (c_2\mathcal{G}_T(t) + c_3)\log\left(\frac{T}{\beta_2^T}\right), \tag{25}$$

where c_1, c_2, c_3 are constants determined by $\beta_1, \beta_2, d, L, \eta$. Then, we will first show that $G_1 \le G$ and suppose that for some $t \in [T]$,

$$G_s \le G, \quad \forall s \in [t] \quad \text{thus} \quad \mathcal{G}_T(s) \le \mathcal{G}_T, \quad \forall s \in [t]. \tag{26}$$

It's then clear to reveal from (25) and the induction assumption that $f(\boldsymbol{y}_{t+1})$ is restricted by the first-order of $\mathcal{G}_T$. Moreover, $f(\boldsymbol{y}_{t+1}) - f^*$ could be served as the upper bound of $\|\bar{\boldsymbol{g}}_{t+1}\|^2$ since

$$\|\bar{\boldsymbol{g}}_{t+1}\|^2 \le \|\nabla f(\boldsymbol{y}_{t+1})\|^2 + \|\bar{\boldsymbol{g}}_{t+1} - \nabla f(\boldsymbol{y}_{t+1})\|^2 \le 2L(f(\boldsymbol{y}_{t+1}) - f^*) + \|\bar{\boldsymbol{g}}_{t+1} - \nabla f(\boldsymbol{y}_{t+1})\|^2, \tag{27}$$

where we use a standard result $\|\nabla f(\boldsymbol{x})\|^2 \le 2L(f(\boldsymbol{x}) - f^*)$ in smooth-based optimization. We also use the smoothness to control $\|\bar{\boldsymbol{g}}_{t+1} - \nabla f(\boldsymbol{y}_{t+1})\|^2$ and combine with (26) and (27) to derive that

$$\|\bar{\boldsymbol{g}}_{t+1}\|^2 \le \tilde{d}_1 + \tilde{d}_2(\sigma_1 G^{p/2} + G),$$

where $\tilde{d}_1, \tilde{d}_2$ are constants that are also determined by hyper-parameters and restricted by $\mathcal{O}(\log T - T\log\beta_2)$ with respect to T. Then, using Young's inequality,

$$\|\bar{\boldsymbol{g}}_{t+1}\|^2 \le \frac{G^2}{2} + \tilde{d}_1 + \frac{4-p}{4}\cdot p^{\frac{p}{4-p}}\left(\tilde{d}_2\right)^{\frac{4}{4-p}} + \left(\tilde{d}_2\right)^2.$$

Thus, combining with a proper construction G^2 (detailed in (52)), we could prove that

$$G^2 = 2\tilde{d}_1 + \frac{4-p}{2}\cdot p^{\frac{p}{4-p}}\left(\tilde{d}_2\right)^{\frac{4}{4-p}} + 2\left(\tilde{d}_2\right)^2,$$

which leads to $\|\bar{\boldsymbol{g}}_{t+1}\|^2 \le G^2$. Combining with the induction argument, we deduce that $\|\bar{\boldsymbol{g}}_t\|^2 \le G^2, \forall t \in [T+1]$.

Final estimation Following the induction step for upper bounding the gradients' norm, we also prove the following result in high probability:

$$L \sum_{s=1}^{T} \frac{\eta_s}{\|\boldsymbol{a}_s\|_\infty} \|\bar{\boldsymbol{g}}_s\|^2 \le L \sum_{s=1}^{T} \eta_s \left\| \frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}} \right\|^2 \le G^2.$$

We could rely on $\mathcal{G}_T$ to prove that $\|\boldsymbol{a}_s\|_\infty \le \mathcal{G}_T \sqrt{1-\beta_2^s} + \epsilon_s, \forall s \in [T]$, and then combine with η_s in Algorithm 1 to further deduce the desired guarantee for $\sum_{s=1}^{T} \|\bar{\boldsymbol{g}}_s\|^2/T$.

7 Conclusion

In this paper, we investigate the convergence of the Adam optimization algorithm on non-convex smooth problems under certain relaxed conditions. We begin by considering a mild noise assumption that encompasses several noise types, particularly the almost surely affine variance noise. Under this noise condition, we demonstrate that Adam can find a stationary point at a rate of $\mathcal{O}(\text{poly}(\log T)/\sqrt{T})$ with high probability. Within our framework, we introduce a novel proxy step-size to manage the entanglement of stochastic gradients and adaptive step-sizes, and we employ a new decomposition method to estimate the errors introduced by the proxy step-size, the momentum, and the corrective terms in Adam.

We also extend our analysis to the convergence of Adam when the objective function is generalized smooth. This relaxed assumption is empirically validated to be more realistic in practical applications. Our results indicate that, with appropriate hyper-parameter tuning, Adam can find a stationary point at the same order of convergence rate as in the smooth case.

Limitations Our study has several limitations that warrant further exploration. First, it would be advantageous to provide experimental results to validate the hyper-parameter settings in our results. Second, the convergence bound is not strictly tight compared to the lower bound, leaving a gap involving logarithmic factors, which may be improved in future work.

Acknowledgement

This work was supported in part by the NSFC under grant number 12471096, and the National Key Research and Development Program of China under grant number 2021YFA1003500.

References

[1] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1-2):165–214, 2023.

[2] Amit Attia and Tomer Koren. SGD with AdaGrad stepsizes: full adaptivity with high probability to unknown parameters, unbounded gradients and affine variance. In *International Conference on Machine Learning*, 2023.

[3] Dimitri P Bertsekas and John N Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

[4] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[5] Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

[6] Jinghui Chen, Dongruo Zhou, Yiqi Tang, Ziyan Yang, Yuan Cao, and Quanquan Gu. Closing the generalization gap of adaptive gradient methods in training deep neural networks. *arXiv preprint arXiv:1806.06763*, 2018.

[7] Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of Adam-type algorithms for non-convex optimization. In *International Conference on Learning Representations*, 2019.

[8] Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized signSGD. In *Advances in Neural Information Processing Systems*, 2022.

[9] Soham De, Anirbit Mukherjee, and Enayat Ullah. Convergence guarantees for RMSProp and Adam in non-convex optimization and an empirical comparison to Nesterov acceleration. *arXiv preprint arXiv:1807.06766*, 2018.

[10] Alexandre Défossez, Leon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of Adam and Adagrad. *Transactions on Machine Learning Research*, 2022.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[12] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7):2121–2159, 2011.

[13] Matthew Faw, Litu Rout, Constantine Caramanis, and Sanjay Shakkottai. Beyond uniform smoothness: a stopped analysis of adaptive SGD. In *Conference on Learning Theory*, 2023.

[14] Matthew Faw, Isidoros Tziotis, Constantine Caramanis, Aryan Mokhtari, Sanjay Shakkottai, and Rachel Ward. The power of adaptivity in SGD: self-tuning step sizes with unbounded gradients and affine variance. In *Conference on Learning Theory*, 2022.

[15] Wayne A Fuller. *Measurement error models*. John Wiley & Sons, 2009.

[16] Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the heavy-ball method for convex optimization. In *European Control Conference*, 2015.

[17] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[18] Zhishuai Guo, Yi Xu, Wotao Yin, Rong Jin, and Tianbao Yang. A novel convergence analysis for algorithms of the Adam family. In *Annual Workshop on Optimization for Machine Learning*, 2021.

[19] Yusu Hong and Junhong Lin. High probability convergence of Adam under unbounded gradients and affine variance noise. *arXiv preprint arXiv:2311.02000*, 2023.

[20] Feihu Huang, Junyi Li, and Heng Huang. Super-Adam: faster and universal framework of adaptive gradients. In *Advances in Neural Information Processing Systems*, 2021.

[21] Fereshte Khani and Percy Liang. Feature noise induces loss discrepancy across groups. In *International Conference on Machine Learning*, pages 5209–5219. PMLR, 2020.

[22] Diederik P Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[23] Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise is not the main factor behind the gap between SGD and Adam on Transformers, but sign descent might be. In *The Eleventh International Conference on Learning Representations*, 2023.

[24] Haochuan Li, Ali Jadbabaie, and Alexander Rakhlin. Convergence of Adam under relaxed assumptions. In *Advances in Neural Information Processing Systems*, 2023.

[25] Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive SGD with momentum. In *Workshop on International Conference on Machine Learning*, 2020.

[26] Mingrui Liu, Youssef Mroueh, Jerret Ross, Wei Zhang, Xiaodong Cui, Payel Das, and Tianbao Yang. Towards better understanding of adaptive gradient algorithms in generative adversarial nets. *arXiv preprint arXiv:1912.11940*, 2019.

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[28] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[29] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[30] Jiang Qian, Yuren Wu, Bojin Zhuang, Shaojun Wang, and Jing Xiao. Understanding gradient clipping in incremental gradient methods. In *International Conference on Artificial Intelligence and Statistics*, 2021.

[31] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018.

[32] Amirhossein Reisizadeh, Haochuan Li, Subhro Das, and Ali Jadbabaie. Variance-reduced clipping for non-convex optimization. *arXiv preprint arXiv:2303.00883*, 2023.

[33] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, pages 400–407, 1951.

[34] Naichen Shi, Dawei Li, Mingyi Hong, and Ruoyu Sun. RMSProp converges with proper hyper-parameter. In *International Conference on Learning Representations*, 2020.

[35] Steve Smale and Yuan Yao. Online learning algorithms. *Foundations of Computational Mathematics*, 6:145–170, 2006.

[36] Matthew Streeter and H Brendan McMahan. Less regret via online conditioning. *arXiv preprint arXiv:1002.4862*, 2010.

[37] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-RMSProp: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 2012.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[39] Bohan Wang, Jingwen Fu, Huishuai Zhang, Nanning Zheng, and Wei Chen. Closing the gap between the upper bound and lower bound of Adam's iteration complexity. In *Advances in Neural Information Processing Systems*, 2023.

[40] Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of AdaGrad for non-convex objectives: simple proofs and relaxed assumptions. In *Conference on Learning Theory*, 2023.

[41] Bohan Wang, Yushun Zhang, Huishuai Zhang, Qi Meng, Zhi-Ming Ma, Tie-Yan Liu, and Wei Chen. Provable adaptivity in Adam. *arXiv preprint arXiv:2208.09900*, 2022.

[42] Ruiqi Wang and Diego Klabjan. Divergence results and convergence of a variance reduced version of adam. *arXiv preprint arXiv:2210.05607*, 2022.

[43] Rachel Ward, Xiaoxia Wu, and Leon Bottou. Adagrad stepsizes: sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, 21(1):9047–9076, 2020.

[44] Shuo Xie and Zhiyuan Li. Implicit bias of adamw: l_∞-norm constrained optimization. In *International Conference on Machine Learning*, 2024.

[45] Huan Xu, Constantine Caramanis, and Shie Mannor. Robust regression and Lasso. In *Advances in Neural Information Processing Systems*, 2008.

[46] Yan Yan, Tianbao Yang, Zhe Li, Qihang Lin, and Yi Yang. A unified analysis of stochastic momentum methods for deep learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018.

[47] Yiming Ying and D-X Zhou. Online regularized classification algorithms. *IEEE Transactions on Information Theory*, 52(11):4775–4788, 2006.

[48] Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Advances in Neural Information Processing Systems*, 2018.

[49] Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved analysis of clipping algorithms for non-convex optimization. In *Advances in Neural Information Processing Systems*, 2020.

[50] Chenyang Zhang, Difan Zou, and Yuan Cao. The implicit bias of Adam on separable data. In *High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*, 2024.

[51] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: a theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020.

[52] Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. Adam can converge without any modification on update rules. In *Advances in Neural Information Processing Systems*, 2022.

[53] Shen-Yi Zhao, Yin-Peng Xie, and Wu-Jun Li. On the convergence and improvement of stochastic normalized gradient descent. *Science China Information Sciences*, 64:1–13, 2021.

[54] Dongruo Zhou, Jinghui Chen, Yuan Cao, Yiqi Tang, Ziyan Yang, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. In *Annual Workshop on Optimization for Machine Learning*, 2020.

[55] Pan Zhou, Xingyu Xie, and Shuicheng Yan. Win: weight-decay-integrated Nesterov acceleration for adaptive gradient algorithms. In *International Conference on Learning Representations*, 2023.

[56] Difan Zou, Yuan Cao, Yuanzhi Li, and Quanquan Gu. Understanding the generalization of adam in learning neural networks with proper regularization. *arXiv preprint arXiv:2108.11371*, 2021.

[57] Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of Adam and RMSProp. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

The appendix is organized as follows. The next section presents some necessary technical lemmas, some of which have appeared in the previous literature. In Appendix B and Appendix C, the detailed proofs for Theorem 3.1 and Theorem 4.1 are presented respectively. Finally, Appendix D and Appendix E provide all the omitted proofs in previous sections.

A Complementary lemmas

We first provide some necessary technical lemmas as follows.

Lemma A.1. *Suppose that $\{\alpha_s\}_{s\ge 1}$ is a non-negative sequence. Given $\beta_2 \in (0,1]$ and $\varepsilon > 0$, we define $\theta_s = \sum_{j=1}^{s} \beta_2^{s-j}\alpha_j$. Then, for any $t \ge 1$,*

$$\sum_{s=1}^{t} \frac{\alpha_j}{\varepsilon + \theta_j} \le \log\left(1 + \frac{\theta_t}{\varepsilon}\right) - t\log\beta_2.$$

Proof. See the proof of Lemma 5.2 in [10]. □

Lemma A.2. *Suppose that $\{\alpha_s\}_{s\ge 1}$ is a real number sequence. Given $0 \le \beta_1 < \beta_2 \le 1$ and $\varepsilon > 0$, we define $\zeta_s = \sum_{j=1}^{s} \beta_1^{s-j}\alpha_j$, $\gamma_s = \frac{1}{1-\beta_1^s}\sum_{j=1}^{s} \beta_1^{s-j}\alpha_j$ and $\theta_s = \sum_{j=1}^{s} \beta_2^{s-j}\alpha_j^2$, then*

$$\sum_{s=1}^{t} \frac{\zeta_s^2}{\varepsilon + \theta_s} \le \frac{1}{(1-\beta_1)(1-\beta_1/\beta_2)}\left(\log\left(1 + \frac{\theta_t}{\varepsilon}\right) - t\log\beta_2\right), \quad \forall t \ge 1,$$
$$\sum_{s=1}^{t} \frac{\gamma_s^2}{\varepsilon + \theta_s} \le \frac{1}{(1-\beta_1)^2(1-\beta_1/\beta_2)}\left(\log\left(1 + \frac{\theta_t}{\varepsilon}\right) - t\log\beta_2\right), \quad \forall t \ge 1.$$

Proof. The proof for the first inequality can be found in the proof of Lemma A.2 [10]. For the second result, let $\hat{M} = \sum_{j=1}^{s} \beta_1^{s-j}$. Then using Jensen's inequality, we have

$$\left(\sum_{j=1}^{s} \beta_1^{s-j}\alpha_j\right)^2 = \left(\hat{M}\sum_{j=1}^{s} \frac{\beta_1^{s-j}}{\hat{M}}\alpha_j\right)^2 \le \hat{M}^2\sum_{j=1}^{s} \frac{\beta_1^{s-j}}{\hat{M}}\alpha_j^2 = \hat{M}\sum_{j=1}^{s} \beta_1^{s-j}\alpha_j^2. \tag{28}$$

Hence, we further have

$$\frac{\gamma_s^2}{\varepsilon + \theta_s} \le \frac{\hat{M}}{(1-\beta_1^s)^2}\sum_{j=1}^{s} \beta_1^{s-j}\frac{\alpha_j^2}{\varepsilon + \theta_s} = \frac{1}{(1-\beta_1)(1-\beta_1^s)}\sum_{j=1}^{s} \beta_1^{s-j}\frac{\alpha_j^2}{\varepsilon + \theta_s}.$$

Recalling the definition of θ_s, we have $\varepsilon + \theta_s \ge \varepsilon + \beta_2^{s-j}\theta_j \ge \beta_2^{s-j}(\varepsilon + \theta_j)$. Hence, combining with $1 - \beta_1 \le 1 - \beta_1^s$,

$$\frac{\gamma_s^2}{\varepsilon + \theta_s} \le \frac{1}{(1-\beta_1)(1-\beta_1^s)}\sum_{j=1}^{s} \left(\frac{\beta_1}{\beta_2}\right)^{s-j}\frac{\alpha_j^2}{\varepsilon + \theta_j} \le \frac{1}{(1-\beta_1)^2}\sum_{j=1}^{s} \left(\frac{\beta_1}{\beta_2}\right)^{s-j}\frac{\alpha_j^2}{\varepsilon + \theta_j}.$$

Summing up both sides over $s \in [t]$, and noting that $\beta_1 < \beta_2$,

$$\sum_{s=1}^{t} \frac{\gamma_s^2}{\varepsilon + \theta_s} \le \frac{1}{(1-\beta_1)^2}\sum_{s=1}^{t}\sum_{j=1}^{s} \left(\frac{\beta_1}{\beta_2}\right)^{s-j}\frac{\alpha_j^2}{\varepsilon + \theta_j} \le \frac{1}{(1-\beta_1)^2(1-\beta_1/\beta_2)}\sum_{j=1}^{t} \frac{\alpha_j^2}{\varepsilon + \theta_j}.$$

Finally applying Lemma A.1, we obtain the desired result. □

Then, we introduce a standard concentration inequality for the martingale difference sequence that is useful for achieving the high probability bounds, see [25] for a proof.

Lemma A.3. *Suppose $\{Z_s\}_{s\in[T]}$ is a martingale difference sequence with respect to $\zeta_1, \cdots, \zeta_T$. Assume that for each $s \in [T]$, σ_s is a random variable only dependent by $\zeta_1, \cdots, \zeta_{s-1}$ and satisfies that*

$$\mathbb{E}\left[\exp(Z_s^2/\sigma_s^2) \mid \zeta_1, \cdots, \zeta_{s-1}\right] \leq \mathrm{e},$$

then for any $\lambda > 0$, and for any $\delta \in (0,1)$, it holds that

$$\mathbb{P}\left(\sum_{s=1}^T Z_s > \frac{1}{\lambda}\log\left(\frac{1}{\delta}\right) + \frac{3}{4}\lambda\sum_{s=1}^T \sigma_s^2\right) \leq \delta.$$

B Proof of Theorem 3.1

The detailed proof of Theorem 3.1 corresponds to the proof sketch in Section 6.

B.1 Preliminary

To start with, we introduce the following two notations,

$$\hat{\boldsymbol{m}}_s = \frac{\boldsymbol{m}_s}{1-\beta_1^s}, \quad \hat{\boldsymbol{v}}_s = \frac{\boldsymbol{v}_s}{1-\beta_2^s}, \tag{29}$$

which include two corrective terms for $\boldsymbol{m}_s$ and $\boldsymbol{v}_s$. It is easy to see that η_s satisfies

$$\eta_s = \frac{\eta\sqrt{1-\beta_2^s}}{1-\beta_1^s} \leq \frac{\eta}{1-\beta_1^s} \leq \frac{\eta}{1-\beta_1}. \tag{30}$$

We follow all the notations in Section 6, which we present here for the convenience of reading,

$$\begin{aligned}
&\mathcal{M}_T = \sqrt{\log\left(\frac{\mathrm{e}T}{\delta}\right)}, \quad G_s = \max_{j\in[s]}\|\bar{\boldsymbol{g}}_j\|, \\
&\mathcal{G}_T(s) = \mathcal{M}_T\sqrt{2\sigma_0^2 + 2\sigma_1^2 G_s^p + 2G_s^2}, \quad \mathcal{G}_T = \mathcal{M}_T\sqrt{2\sigma_0^2 + 2\sigma_1^2 G^p + 2G^2}, \\
&\boldsymbol{b}_s = \sqrt{\beta_2\boldsymbol{v}_{s-1} + (1-\beta_2)\boldsymbol{g}_s^2} + \boldsymbol{\epsilon}_s, \\
&\boldsymbol{a}_s = \sqrt{\beta_2\boldsymbol{v}_{s-1} + (1-\beta_2)\left(\mathcal{G}_T(s)\mathbf{1}_d\right)^2} + \boldsymbol{\epsilon}_s.
\end{aligned}$$

The following lemmas provide some estimations for the algorithm-dependent terms, which play vital roles in the proof of Theorem 3.1. The detailed proofs could be found in Appendix D.1.

Lemma B.1. *Let $\eta_s, \boldsymbol{b}_s$ be given in Algorithm 1 and* (16)*, then*

$$\left\|\frac{\eta_s\boldsymbol{b}_{s-1}}{\eta_{s-1}\boldsymbol{b}_s} - \mathbf{1}_d\right\|_\infty \leq \Sigma_{\max} := \max\left\{1, \sqrt{\frac{1+\beta_2}{\beta_2}} - 1\right\}, \quad \forall s \geq 2.$$

The following lemma could be found similarly in [56, Lemma A.2].

Lemma B.2. *Let $\boldsymbol{m}_s, \boldsymbol{b}_s$ be given in Algorithm 1 and* (16) *with $0 \leq \beta_1 < \beta_2 < 1$, respectively. Then,*

$$\left\|\frac{\boldsymbol{m}_s}{\boldsymbol{b}_s}\right\|_\infty \leq \sqrt{\frac{(1-\beta_1)(1-\beta_1^s)}{(1-\beta_2)(1-\beta_1/\beta_2)}}, \quad \forall s \geq 1.$$

Consequently, if f is L-smooth and we set $\eta = C_0\sqrt{1-\beta_2}$ for some constant $C_0 > 0$, then

$$\|\bar{\boldsymbol{g}}_s\| \leq \|\bar{\boldsymbol{g}}_1\| + LC_0 s\sqrt{\frac{d}{1-\beta_1/\beta_2}}, \quad \forall s \geq 1.$$

The following lemma is necessary for deriving (24) in the proof sketch.

Lemma B.3. *Let $\boldsymbol{g}_s, \boldsymbol{m}_s$ be given in Algorithm 1 and $\hat{\boldsymbol{m}}_s, \boldsymbol{b}_s$ be defined in* (29) *and* (16). *If* $0 \le \beta_1 < \beta_2 < 1$ *and* $\mathcal{F}_i(t) = 1 + \frac{1}{\epsilon^2}\sum_{s=1}^t g_{s,i}^2$*, then for any* $t \ge 1$,

$$\sum_{s=1}^t \left\|\frac{\boldsymbol{g}_s}{\boldsymbol{b}_s}\right\|^2 \le \frac{1}{1-\beta_2}\sum_{i=1}^d \log\left(\frac{\mathcal{F}_i(t)}{\beta_2^t}\right),$$

$$\sum_{s=1}^t \left\|\frac{\boldsymbol{m}_s}{\boldsymbol{b}_s}\right\|^2 \le \frac{1-\beta_1}{(1-\beta_2)(1-\beta_1/\beta_2)}\sum_{i=1}^d \log\left(\frac{\mathcal{F}_i(t)}{\beta_2^t}\right),$$

$$\sum_{s=1}^t \left\|\frac{\boldsymbol{m}_s}{\boldsymbol{b}_{s+1}}\right\|^2 \le \frac{1-\beta_1}{\beta_2(1-\beta_2)(1-\beta_1/\beta_2)}\sum_{i=1}^d \log\left(\frac{\mathcal{F}_i(t)}{\beta_2^t}\right),$$

$$\sum_{s=1}^t \left\|\frac{\hat{\boldsymbol{m}}_s}{\boldsymbol{b}_s}\right\| \le \frac{1}{(1-\beta_2)(1-\beta_1/\beta_2)}\sum_{i=1}^d \log\left(\frac{\mathcal{F}_i(t)}{\beta_2^t}\right).$$

The following lemmas are based on the smooth condition.

Lemma B.4. *Suppose that f is L-smooth and Assumption (A1) holds, then for any $\boldsymbol{x} \in \mathbb{R}^d$,*

$$\|\nabla f(\boldsymbol{x})\|^2 \le 2L(f(\boldsymbol{x}) - f^*).$$

Lemma B.5. *Let $\boldsymbol{x}_s$ be given in Algorithm 1 and $\boldsymbol{y}_s$ be defined in* (17). *If f is L-smooth, $\eta = C_0\sqrt{1-\beta_2}$ and $0 \le \beta_1 < \beta_2 < 1$, then*

$$\|\nabla f(\boldsymbol{x}_s)\| \le \|\nabla f(\boldsymbol{y}_s)\| + M, \quad M := \frac{LC_0\sqrt{d}}{(1-\beta_1)\sqrt{1-\beta_1/\beta_2}}, \quad \forall s \ge 1.$$

B.2 Start point and decomposition

Specifically, we fix the horizon T and start from the descent lemma of L-smoothness,

$$f(\boldsymbol{y}_{s+1}) \le f(\boldsymbol{y}_s) + \langle \nabla f(\boldsymbol{y}_s), \boldsymbol{y}_{s+1} - \boldsymbol{y}_s\rangle + \frac{L}{2}\|\boldsymbol{y}_{s+1} - \boldsymbol{y}_s\|^2, \quad \forall s \in [T]. \tag{31}$$

For any given $t \in [T]$, combining with (18) and (31) and then summing over $s \in [t]$, we obtain the same inequality in (20),

$$\begin{aligned} f(\boldsymbol{y}_{t+1}) \le f(\boldsymbol{x}_1) + \underbrace{\sum_{s=1}^t -\eta_s \left\langle \nabla f(\boldsymbol{y}_s), \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s}\right\rangle}_{\textbf{A}} + \underbrace{\frac{\beta_1}{1-\beta_1}\sum_{s=1}^t \langle \Delta_s \odot (\boldsymbol{x}_s - \boldsymbol{x}_{s-1}), \nabla f(\boldsymbol{y}_s)\rangle}_{\textbf{B}} \\ + \underbrace{\frac{L}{2}\sum_{s=1}^t \left\|\eta_s \cdot \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s} - \frac{\beta_1}{1-\beta_1}(\Delta_s \odot (\boldsymbol{x}_s - \boldsymbol{x}_{s-1}))\right\|^2}_{\textbf{C}}, \end{aligned} \tag{32}$$

where we use Δ_s in (20) and $\boldsymbol{y}_1 = \boldsymbol{x}_1$. We then further make a decomposition by introducing $\bar{\boldsymbol{g}}_s$ into **A** and **B**

$$\textbf{A} = \underbrace{\sum_{s=1}^t -\eta_s \left\langle \bar{\boldsymbol{g}}_s, \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s}\right\rangle}_{\textbf{A.1}} + \underbrace{\sum_{s=1}^t \eta_s \left\langle \bar{\boldsymbol{g}}_s - \nabla f(\boldsymbol{y}_s), \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s}\right\rangle}_{\textbf{A.2}}, \tag{33}$$

and

$$\textbf{B} = \underbrace{\frac{\beta_1}{1-\beta_1}\sum_{s=1}^t \langle \Delta_s \odot (\boldsymbol{x}_s - \boldsymbol{x}_{s-1}), \bar{\boldsymbol{g}}_s\rangle}_{\textbf{B.1}} + \underbrace{\frac{\beta_1}{1-\beta_1}\sum_{s=1}^t \langle \Delta_s \odot (\boldsymbol{x}_s - \boldsymbol{x}_{s-1}), \nabla f(\boldsymbol{y}_s) - \bar{\boldsymbol{g}}_s\rangle}_{\textbf{B.2}}. \tag{34}$$

B.3 Probabilistic estimations

We will provide two probabilistic inequalities with the detailed proofs given in Appendix D.2. The first one establishes an upper bound for the noise norm, which we have already informally presented in (21).

Lemma B.6. *Given $T \geq 1$, suppose that for any $s \in [T]$, $\boldsymbol{\xi}_s = \boldsymbol{g}_s - \bar{\boldsymbol{g}}_s$ satisfies Assumption (A3). Then for any given $\delta \in (0,1)$, it holds that with probability at least $1-\delta$,*

$$\|\boldsymbol{\xi}_s\|^2 \leq \mathcal{M}_T^2 \left(\sigma_0^2 + \sigma_1^2 \|\bar{\boldsymbol{g}}_s\|^p\right), \quad \forall s \in [T]. \tag{35}$$

We next provide a probabilistic upper bound as shown in (22) for a summation of the inner product, where we rely on the property of the martingale difference sequence and the proxy step-size $\boldsymbol{a}_s$ in (23).

Lemma B.7. *Given $T \geq 1$ and $\delta \in (0,1)$. If Assumptions (A2) and (A3) hold, then for any $\lambda > 0$, with probability at least $1-\delta$,*

$$-\sum_{s=1}^{t} \eta_s \left\langle \bar{\boldsymbol{g}}_s, \frac{\boldsymbol{\xi}_s}{\boldsymbol{a}_s} \right\rangle \leq \frac{3\lambda\eta\mathcal{G}_T(t)}{4(1-\beta_1)\sqrt{1-\beta_2}} \sum_{s=1}^{t} \eta_s \left\| \frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}} \right\|^2 + \frac{d}{\lambda} \log\left(\frac{dT}{\delta}\right), \quad \forall t \in [T]. \tag{36}$$

As a consequence, when setting $\lambda = (1-\beta_1)\sqrt{1-\beta_2}/(3\eta\mathcal{G}_T)$, it holds that with probability at least $1-\delta$,

$$-\sum_{s=1}^{t} \eta_s \left\langle \bar{\boldsymbol{g}}_s, \frac{\boldsymbol{\xi}_s}{\boldsymbol{a}_s} \right\rangle \leq \frac{\mathcal{G}_T(t)}{4\mathcal{G}_T} \sum_{s=1}^{t} \eta_s \left\| \frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}} \right\|^2 + D_1 \mathcal{G}_T, \quad \forall t \in [T], \tag{37}$$

where $D_1 = \frac{3\eta}{(1-\beta_1)\sqrt{1-\beta_2}} \log\left(\frac{T}{\delta}\right)$.

B.4 Deterministic estimations

In this section, we shall assume that (35) or/and (37) hold whenever the related estimation is needed. Then we obtain the following key lemmas with the detailed proofs given in Appendix D.3.

Lemma B.8. *Given $T \geq 1$. If (35) holds, then we have*

$$\max_{j\in[s]} \|\boldsymbol{\xi}_j\| \leq \mathcal{G}_T(s), \quad \max_{j\in[s]} \|\boldsymbol{g}_j\| \leq \mathcal{G}_T(s), \quad \max_{j\in[s]} \|\boldsymbol{v}_j\|_\infty \leq (\mathcal{G}_T(s))^2, \quad \forall s \in [T].$$

Lemma B.9. *Given $T \geq 1$. If $\mathbf{b}_s = (b_{s,i})_i$ and $\boldsymbol{a}_s = (a_{s,i})_i$ follow the definitions in (16) and (23) respectively, and (35) holds, then for all $s \in [T], i \in [d]$,*

$$\left| \frac{1}{a_{s,i}} - \frac{1}{b_{s,i}} \right| \leq \frac{\mathcal{G}_T(s)\sqrt{1-\beta_2}}{a_{s,i} b_{s,i}} \quad \text{and} \quad \left| \frac{1}{a_{s,i}} - \frac{1}{b_{s-1,i}} \right| \leq \frac{(\mathcal{G}_T(s)+\epsilon)\sqrt{1-\beta_2}}{a_{s,i} b_{s-1,i}}.$$

Lemma B.10. *Given $T \geq 1$. Under the conditions in Lemma B.3 and Lemma B.5, if (35) holds, then the following inequality holds,*

$$\mathcal{F}_i(t) \leq \mathcal{F}(T), \quad \forall t \in [T], i \in [d],$$

where $\hat{M} = M(1-\beta_1)$ and M follows the definition in Lemma B.5, $\mathcal{F}(T)$ is define by

$$\mathcal{F}(T) := 1 + \frac{2\mathcal{M}_T^2}{\epsilon^2} \left[\sigma_0^2 T + \sigma_1^2 T \left(\|\bar{\boldsymbol{g}}_1\| + T\hat{M} \right)^p + T \left(\|\bar{\boldsymbol{g}}_1\| + T\hat{M} \right)^2 \right]. \tag{38}$$

We move to estimate all the related terms in Appendix B.2. First, the estimation for **A.1** relies on both the two probabilistic estimations in Appendix B.3.

Lemma B.11. *Given $T \geq 1$, suppose that (35) and (37) hold. Then for all $t \in [T]$,*

$$\mathbf{A.1} \leq \left(\frac{\mathcal{G}_T(t)}{4\mathcal{G}} - \frac{3}{4} \right) \sum_{s=1}^{t} \eta_s \left\| \frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}} \right\|^2 + D_1 \mathcal{G}_T + D_2 \mathcal{G}_T(t) \sum_{s=1}^{t} \left\| \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s} \right\|^2, \tag{39}$$

where D_1 is given as in Lemma B.7 and $D_2 = \frac{\eta\sqrt{1-\beta_2}}{1-\beta_1}$.

We also obtain the following lemma to estimate the adaptive momentum part **B.1**.

Lemma B.12. *Given $T \geq 1$, if* (35) *holds, then for all $t \in [T]$,*

$$\mathbf{B.1} \leq \frac{1}{4}\sum_{s=1}^{t} \eta_s \left\| \frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}} \right\|^2 + (D_3 \mathcal{G}_T(t) + D_4) \sum_{s=1}^{t} \left(\left\| \frac{\boldsymbol{m}_{s-1}}{\boldsymbol{b}_s} \right\|^2 + \left\| \frac{\boldsymbol{m}_{s-1}}{\boldsymbol{b}_{s-1}} \right\|^2 \right) + D_5 G_t, \quad (40)$$

where

$$D_3 = \frac{2\eta\sqrt{1-\beta_2}}{(1-\beta_1)^3}, \quad D_4 = \epsilon D_3, \quad D_5 = \frac{2\eta\sqrt{d}}{\sqrt{(1-\beta_1)^3(1-\beta_2)(1-\beta_1/\beta_2)}}. \quad (41)$$

Proposition B.13. *Given $T \geq 1$. If f is L-smooth, then the following inequality holds,*

$$f(\boldsymbol{y}_{t+1}) \leq f(\boldsymbol{x}_1) + \mathbf{A.1} + \mathbf{B.1} + D_6 \sum_{s=1}^{t-1} \left\| \frac{\hat{\boldsymbol{m}}_s}{\boldsymbol{b}_s} \right\|^2 + D_7 \sum_{s=1}^{t} \left\| \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s} \right\|^2, \quad \forall t \in [T],$$

where $\Sigma_{\max}$ is as in Lemma B.1 and

$$D_6 = \frac{L\eta^2(1+4\Sigma_{\max}^2)}{2(1-\beta_1)^2}, \quad D_7 = \frac{3L\eta^2}{2(1-\beta_1)^2}. \quad (42)$$

Proof. Recalling the decomposition in Appendix B.2. We first estimate **A.2**. Using the smoothness of f and (17), we have

$$\|\nabla f(\boldsymbol{y}_s) - \bar{\boldsymbol{g}}_s\| \leq L\|\boldsymbol{y}_s - \boldsymbol{x}_s\| = \frac{L\beta_1}{1-\beta_1}\|\boldsymbol{x}_s - \boldsymbol{x}_{s-1}\|. \quad (43)$$

Hence, applying Young's inequality, (43) and (30),

$$\begin{aligned} &\eta_s \left\langle \bar{\boldsymbol{g}}_s - \nabla f(\boldsymbol{y}_s), \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s} \right\rangle \leq \eta_s \|\bar{\boldsymbol{g}}_s - \nabla f(\boldsymbol{y}_s)\| \cdot \left\| \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s} \right\| \\ \leq\ &\frac{1}{2L}\|\bar{\boldsymbol{g}}_s - \nabla f(\boldsymbol{y}_s)\|^2 + \frac{L\eta_s^2}{2}\left\| \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s} \right\|^2 \leq \frac{L\beta_1^2}{2(1-\beta_1)^2}\|\boldsymbol{x}_s - \boldsymbol{x}_{s-1}\|^2 + \frac{L\eta^2}{2(1-\beta_1)^2}\left\| \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s} \right\|^2. \quad (44) \end{aligned}$$

Recalling the updated rule in Algorithm 1 and applying (29) as well as (30),

$$\|\boldsymbol{x}_s - \boldsymbol{x}_{s-1}\|^2 = \eta_{s-1}^2 \left\| \frac{\boldsymbol{m}_{s-1}}{\boldsymbol{b}_{s-1}} \right\|^2 \leq \eta^2 \left\| \frac{\hat{\boldsymbol{m}}_{s-1}}{\boldsymbol{b}_{s-1}} \right\|^2. \quad (45)$$

Therefore, applying (44), (45) and $\beta_1 \in [0, 1)$, and then summing over $s \in [t]$

$$\mathbf{A.2} \leq \frac{L\eta^2}{2(1-\beta_1)^2}\sum_{s=1}^{t} \left\| \frac{\hat{\boldsymbol{m}}_{s-1}}{\boldsymbol{b}_{s-1}} \right\|^2 + \frac{L\eta^2}{2(1-\beta_1)^2}\sum_{s=1}^{t} \left\| \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s} \right\|^2. \quad (46)$$

Applying Cauchy-Schwarz inequality, Lemma B.1, and combining with (43), (45), $\Sigma_{\max} \geq 1$, and $\beta_1 \in [0, 1)$

$$\begin{aligned} \mathbf{B.2}\ &\leq \frac{\beta_1}{1-\beta_1}\sum_{s=1}^{t} \|\Delta_s\|_\infty \|\boldsymbol{x}_s - \boldsymbol{x}_{s-1}\| \|\nabla f(\boldsymbol{y}_s) - \bar{\boldsymbol{g}}_s\| \\ &\leq \frac{L\beta_1^2\Sigma_{\max}}{(1-\beta_1)^2}\sum_{s=1}^{t} \|\boldsymbol{x}_s - \boldsymbol{x}_{s-1}\|^2 \leq \frac{L\Sigma_{\max}^2\eta^2}{(1-\beta_1)^2}\sum_{s=1}^{t} \left\| \frac{\hat{\boldsymbol{m}}_{s-1}}{\boldsymbol{b}_{s-1}} \right\|^2, \quad (47) \end{aligned}$$

Finally, applying the basic inequality, Lemma B.1 and (45),

$$\begin{aligned} \mathbf{C} &\leq L\sum_{s=1}^{t} \eta_s^2 \left\| \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s} \right\|^2 + \frac{L\beta_1^2}{(1-\beta_1)^2}\sum_{s=1}^{t} \|\Delta_s\|_\infty^2 \|\boldsymbol{x}_s - \boldsymbol{x}_{s-1}\|^2 \\ &\leq \frac{L\eta^2}{(1-\beta_1)^2}\sum_{s=1}^{t} \left\| \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s} \right\|^2 + \frac{L\eta^2\Sigma_{\max}^2}{(1-\beta_1)^2}\sum_{s=1}^{t} \left\| \frac{\hat{\boldsymbol{m}}_{s-1}}{\boldsymbol{b}_{s-1}} \right\|^2. \quad (48) \end{aligned}$$

Recalling the decomposition in (33) and (34), then plugging (46), (47) and (48) into (32), we obtain the desired result. □

B.5 Bounding gradients

Based on all the results in Appendix B.3 and Appendix B.4, we are now ready to provide a global upper bound for gradients' norm along the optimization trajectory.

Proposition B.14. *Under the same conditions in Theorem 3.1, for any given $\delta \in (0, 1/2)$, it holds that with probability at least $1 - 2\delta$,*

$$\|\bar{\boldsymbol{g}}_t\|^2 \le G_t^2 \le G^2, \quad \|\boldsymbol{g}_t\|^2 \le (\mathcal{G}_T(t))^2 \le \mathcal{G}_T^2, \quad \forall t \in [T+1], \tag{49}$$

and

$$\|\bar{\boldsymbol{g}}_{t+1}\|^2 \le G^2 - L\sum_{s=1}^{t}\eta_s \left\|\frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}}\right\|^2, \quad \forall t \in [T], \tag{50}$$

where G^2 is as in Theorem 3.1 and $G_t, G, \mathcal{G}_T$ are given by (19).

Proof. Applying Lemma B.6 and Lemma B.7, we know that (35) or (37) hold with probability at least $1 - \delta$. With these two inequalities, we could deduce the desired inequalities (49) and (50). Therefore, (49) and (50) hold with probability at least $1 - 2\delta$. We first plug (39) and (40) into the result in Proposition B.13, which leads to that for all $t \in [T]$,

$$\begin{aligned} f(\boldsymbol{y}_{t+1}) \le f(\boldsymbol{x}_1) &+ \left(\frac{\mathcal{G}_T(t)}{4\mathcal{G}_T} - \frac{1}{2}\right)\sum_{s=1}^{t}\eta_s\left\|\frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}}\right\|^2 + D_1\mathcal{G}_T + (D_2\mathcal{G}_T(t) + D_7)\sum_{s=1}^{t}\left\|\frac{\boldsymbol{g}_s}{\boldsymbol{b}_s}\right\|^2 \\ &+ (D_3\mathcal{G}_T(t) + D_4)\sum_{s=1}^{t}\left(\left\|\frac{\boldsymbol{m}_{s-1}}{\boldsymbol{b}_s}\right\|^2 + \left\|\frac{\boldsymbol{m}_{s-1}}{\boldsymbol{b}_{s-1}}\right\|^2\right) + D_5 G_t + D_6\sum_{s=1}^{t-1}\left\|\frac{\hat{\boldsymbol{m}}_s}{\boldsymbol{b}_s}\right\|^2. \end{aligned} \tag{51}$$

Next, we will introduce the induction argument based on (51). We first provide the specific definition of G^2 as follows which is a constant determined by the horizon T and other hyper-parameters but not relying on t,[4]

$$\begin{aligned} G^2 := 8L(f(\boldsymbol{x}_1) - f^*) &+ \frac{48\mathcal{M}_T L C_0\sigma_0}{1-\beta_1}\log\left(\frac{T}{\delta}\right) + \frac{16\mathcal{M}_T L C_0\sigma_0 d}{1-\beta_1}\log\left(\frac{\mathcal{F}(T)}{\beta_2^T}\right) \\ &+ 8\left(\frac{3LC_0 + 8(\mathcal{M}_T\sigma_0 + \epsilon_0)}{\beta_2}\right)\frac{LC_0 d}{(1-\beta_1)^2(1-\beta_1/\beta_2)}\log\left(\frac{\mathcal{F}(T)}{\beta_2^T}\right) \\ &+ \frac{4-p}{2}\cdot p^{\frac{p}{4-p}}\left[\frac{72\mathcal{M}_T L\sigma_1 C_0 d}{\beta_2(1-\beta_1)^2(1-\beta_1/\beta_2)}\log\left(\frac{T+\mathcal{F}(T)}{\delta\beta_2^T}\right)\right]^{\frac{4}{4-p}} \\ &+ 32\left[\frac{18\mathcal{M}_T L C_0 d}{\beta_2(1-\beta_1)^2(1-\beta_1/\beta_2)}\log\left(\frac{T+\mathcal{F}(T)}{\delta\beta_2^T}\right)\right]^2 + \frac{4L^2C_0^2 d}{(1-\beta_1)^2(1-\beta_1/\beta_2)}. \end{aligned} \tag{52}$$

The induction then begins by noting that $G_1^2 = \|\bar{\boldsymbol{g}}_1\|^2 \le 2L(f(\boldsymbol{x}_1) - f^*) \le G^2$ from Lemma B.4 and (52). Then we assume that for some $t \in [T]$,

$$G_s \le G, \quad \forall s \in [t] \quad \text{consequently} \quad \mathcal{G}_T(s) \le \mathcal{G}_T, \quad \forall s \in [t]. \tag{53}$$

Using this induction assumption over (51) and subtracting with f^* on both sides,

$$\begin{aligned} f(\boldsymbol{y}_{t+1}) - f^* \le f(\boldsymbol{x}_1) - f^* &- \frac{1}{4}\sum_{s=1}^{t}\eta_s\left\|\frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}}\right\|^2 + D_1\mathcal{G}_T + (D_2\mathcal{G}_T + D_7)\sum_{s=1}^{t}\left\|\frac{\boldsymbol{g}_s}{\boldsymbol{b}_s}\right\|^2 \\ &+ (D_3\mathcal{G}_T + D_4)\sum_{s=1}^{t}\left(\left\|\frac{\boldsymbol{m}_{s-1}}{\boldsymbol{b}_s}\right\|^2 + \left\|\frac{\boldsymbol{m}_{s-1}}{\boldsymbol{b}_{s-1}}\right\|^2\right) + D_5 G + D_6\sum_{s=1}^{t-1}\left\|\frac{\hat{\boldsymbol{m}}_s}{\boldsymbol{b}_s}\right\|^2. \end{aligned} \tag{54}$$

Further, we combine with Lemma B.3 and Lemma B.10 to estimate the four summations defined in Lemma B.3, and then use $G \le \mathcal{G}_T \le 2\mathcal{M}_T\left(\sigma_0 + \sigma_1 G^{p/2} + G\right)$ to control the RHS of (54),

$$f(\boldsymbol{y}_{t+1}) - f^* \le -\frac{1}{4}\sum_{s=1}^{t}\eta_s\left\|\frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}}\right\|^2 + \tilde{D}_1 + \tilde{D}_2 + \tilde{D}_3\mathcal{H}(G), \tag{55}$$

[4] We further deduce (7) in Theorem 3.1 based on (52).

where $\mathcal{H}(G) = \sigma_1 G^{p/2} + G$ and $\tilde{D}_1, \tilde{D}_2, \tilde{D}_3$ are defined as

$$\tilde{D}_1 = f(\boldsymbol{x}_1) - f^* + 2\mathcal{M}_T \sigma_0 D_1,$$
$$\tilde{D}_2 = \left[\frac{2\mathcal{M}_T\sigma_0 D_2 + D_7}{1-\beta_2} + \frac{4\left(\mathcal{M}_T\sigma_0 D_3 + D_4\right)(1-\beta_1)}{\beta_2(1-\beta_2)(1-\beta_1/\beta_2)} + \frac{D_6}{(1-\beta_2)(1-\beta_1/\beta_2)}\right] d\log\left(\frac{\mathcal{F}(T)}{\beta_2^T}\right),$$
$$\tilde{D}_3 = 2\mathcal{M}_T\left[D_1 + \left(\frac{D_2 d}{1-\beta_2} + \frac{2D_3(1-\beta_1)d}{\beta_2(1-\beta_2)(1-\beta_1/\beta_2)}\right)\log\left(\frac{\mathcal{F}(T)}{\beta_2^T}\right)\right] + D_5.$$

Applying Lemma B.5 and Lemma B.4,

$$\|\bar{\boldsymbol{g}}_{t+1}\|^2 \le 2\|\nabla f(\boldsymbol{y}_{t+1})\|^2 + 2M^2 \le 4L(f(\boldsymbol{y}_{t+1}) - f^*) + 2M^2. \tag{56}$$

Then combining (55) with (56),

$$\|\bar{\boldsymbol{g}}_{t+1}\|^2 \le -L\sum_{s=1}^{t}\eta_s\left\|\frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}}\right\|^2 + 4L(\tilde{D}_1 + \tilde{D}_2) + 4L\tilde{D}_3\mathcal{H}(G) + 2M^2.$$

Applying two Young's inequalities where $ab \le \frac{a^2}{2} + \frac{b^2}{2}$ and $ab^{\frac{p}{2}} \le \frac{4-p}{4}\cdot a^{\frac{4}{4-p}} + \frac{p}{4}\cdot b^2, \forall a, b \ge 0$,

$$\begin{aligned}\|\bar{\boldsymbol{g}}_{t+1}\|^2 \le \frac{G^2}{4} + \frac{G^2}{4} - L\sum_{s=1}^{t}\eta_s\left\|\frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}}\right\|^2 + 4L(\tilde{D}_1 + \tilde{D}_2)\\ + 16L^2\tilde{D}_3^2 + \frac{4-p}{4}\cdot p^{\frac{p}{4-p}}\left(4L\sigma_1\tilde{D}_3\right)^{\frac{4}{4-p}} + 2M^2.\end{aligned} \tag{57}$$

Recalling the definitions of $D_i, i \in [7]$ in (37), (39), (41), and (42). With a simple calculation relying on $\eta = C_0\sqrt{1-\beta_2}, \epsilon \le \epsilon_0, \Sigma_{\max} \le 1/\sqrt{\beta_2}$ and $0 \le \beta_1 < \beta_2 < 1$, we could deduce that G^2 given in (52) satisfies

$$G^2 = 8L(\tilde{D}_1 + \tilde{D}_2) + 32L^2\tilde{D}_3^2 + \frac{4-p}{2}\cdot p^{\frac{p}{4-p}}\left(4L\sigma_1\tilde{D}_3\right)^{\frac{4}{4-p}} + 4M^2. \tag{58}$$

Based on (57) and (58), we then deduce that $\|\bar{\boldsymbol{g}}_{t+1}\|^2 \le G^2$. Further combining with G_{t+1} in (19) and the induction assumption in (53),

$$G_{t+1} \le \max\{\|\bar{\boldsymbol{g}}_{t+1}\|, G_t\} \le G.$$

Hence, the induction is complete and we obtain the desired result in (49). Furthermore, as a consequence of (57), we also prove that (50) holds. □

B.6 Proof of the main result

Now we are ready to prove the main convergence result.

Proof of Theorem 3.1. We set $t = T$ in (50) to obtain that with probability at least $1 - 2\delta$,

$$L\sum_{s=1}^{T}\frac{\eta_s}{\|\boldsymbol{a}_s\|_\infty}\|\bar{\boldsymbol{g}}_s\|^2 \le L\sum_{s=1}^{T}\eta_s\left\|\frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}}\right\|^2 \le G^2 - \|\bar{\boldsymbol{g}}_{T+1}\|^2 \le G^2. \tag{59}$$

Then, in what follows, we will assume that both (49) and (59) hold. Based on these two inequalities, we could derive the final convergence bound. Since (49) and (59) hold with probability at least $1-2\delta$, the final convergence bound also holds with probability at least $1-2\delta$. Applying $\boldsymbol{a}_s$ in (23) and (49), we have

$$\begin{aligned}\|\boldsymbol{a}_s\|_\infty &= \max_{i\in[d]}\sqrt{\beta_2 v_{s-1,i} + (1-\beta_2)(\mathcal{G}_T(s))^2} + \epsilon_s\\ &\le \max_{i\in[d]}\sqrt{(1-\beta_2)\left[\sum_{j=1}^{s-1}\beta_2^{s-j}g_{j,i}^2 + (\mathcal{G}_T(s))^2\right]} + \epsilon_s\\ &\le \sqrt{(1-\beta_2)\sum_{j=1}^{s}\beta_2^{s-j}\mathcal{G}_T^2} + \epsilon_s = \mathcal{G}_T\sqrt{1-\beta_2^s} + \epsilon_s, \quad \forall s \in [T].\end{aligned} \tag{60}$$

Then combining with the setting η_s and ϵ_s in (6), we have for any $s \in [T]$,

$$\frac{\eta_s}{\|\boldsymbol{a}_s\|_\infty} \geq \frac{C_0\sqrt{(1-\beta_2^s)(1-\beta_2)}}{\mathcal{G}_T\sqrt{1-\beta_2^s}+\epsilon_0\sqrt{(1-\beta_2^s)(1-\beta_2)}} \cdot \frac{1}{1-\beta_1^s} \geq \frac{C_0\sqrt{1-\beta_2}}{\mathcal{G}_T+\epsilon_0\sqrt{1-\beta_2}}.$$

We therefore combine with (59) to obtain that with probability at least $1-2\delta$,

$$\frac{1}{T}\sum_{s=1}^T \|\bar{\boldsymbol{g}}_s\|^2 \leq \frac{G^2}{TLC_0}\left(\frac{\sqrt{2\sigma_0^2+2\sigma_1^2G^p+2G^2}}{\sqrt{1-\beta_2}}+\epsilon_0\right)\sqrt{\log\left(\frac{\mathrm{e}T}{\delta}\right)}. \tag{61}$$

Since $\beta_2 \in (0,1)$, we have

$$-\log\beta_2 = \log\left(\frac{1}{\beta_2}\right) \leq \frac{1-\beta_2}{\beta_2} = \frac{c}{T\beta_2},$$

where we apply $\log(1/a) \leq (1-a)/a, \forall a \in (0,1)$. With both sides multiplying T, we obtain that $\log\left(1/\beta_2^T\right) \leq c/\beta_2$. Then, we further have that when $\beta_2 = 1 - c/T$,

$$\log\left(\frac{T}{\beta_2^T}\right) \leq \log T + \frac{c}{\beta_2}. \tag{62}$$

Since $0 \leq \beta_1 < \beta_2 < 1$, there exists some constants $\varepsilon_1, \varepsilon_2 > 0$ such that

$$\frac{1}{\beta_2} \leq \frac{1}{\varepsilon_1}, \quad \frac{1}{1-\beta_1/\beta_2} \leq \frac{1}{\varepsilon_2}. \tag{63}$$

Therefore combining (62), (63) and (52), we could verify that $G^2 \sim \mathcal{O}\left(\text{poly}(\log T)\right)$ with respect to T. Finally, using the convergence result in (61), we obtain the desired result. □

C Proof of Theorem 4.1

In this section, we shall follow all the notations defined in Section 6. Further, we will add two non-decreasing sequences $\{\mathcal{L}_s^{(x)}\}_{s\geq 1}$ and $\{\mathcal{L}_s^{(y)}\}_{s\geq 1}$ as follows

$$\mathcal{L}_s^{(x)} = L_0 + L_q G_s^q, \quad \mathcal{L}_s^{(y)} = L_0 + L_q(G_s + G_s^q + L_0/L_q)^q, \quad \forall s \geq 1. \tag{64}$$

C.1 Preliminary

We first mention that Lemma B.1, Lemma B.2, and Lemma B.3 in Appendix B.1 remain unchanged since they are independent of the smooth condition. Then the first essential challenge is that we need to properly tune η to restrict the distance between $\boldsymbol{x}_{s+1}$ and $\boldsymbol{x}_s$, $\boldsymbol{y}_{s+1}$ and $\boldsymbol{y}_s$ within $1/L_q$ for all $s \geq 1$. The following two lemmas then ensure this point. The detailed proofs could be found in Appendix E.

Lemma C.1. *Let $\boldsymbol{x}_s, \boldsymbol{y}_s$ be defined in Algorithm 1 and* (17)*. If $0 \leq \beta_1 < \beta_2 < 1$, then for any $s \geq 1$,*

$$\max\{\|\boldsymbol{x}_{s+1}-\boldsymbol{x}_s\|, \|\boldsymbol{y}_s-\boldsymbol{x}_s\|, \|\boldsymbol{y}_{s+1}-\boldsymbol{y}_s\|\} \leq \eta\sqrt{\frac{4d}{\beta_2(1-\beta_1)^2(1-\beta_2)(1-\beta_1/\beta_2)}}. \tag{65}$$

As a consequence, when

$$\eta \leq \frac{1}{L_q F}, \quad F := \sqrt{\frac{4d}{\beta_2(1-\beta_1)^2(1-\beta_2)(1-\beta_1/\beta_2)}}, \tag{66}$$

then for any $s \geq 1$, all the three gaps in (65) *are smaller than $1/L_q$.*

Lemma C.2. *Let $\eta \leq 1/(L_q F)$ where F is as in Lemma C.1. If f is (L_0, L_q)-smooth, then for any $s \geq 1$,*

$$\begin{aligned}\|\nabla f(\boldsymbol{y}_s)\| &\leq L_0/L_q + \|\nabla f(\boldsymbol{x}_s)\|^q + \|\nabla f(\boldsymbol{x}_s)\|,\\ \|\nabla f(\boldsymbol{x}_s)\| &\leq L_0/L_q + \|\nabla f(\boldsymbol{y}_s)\|^q + \|\nabla f(\boldsymbol{y}_s)\|.\end{aligned}$$

As a consequence, for any $s \geq 1$,

$$\|\nabla f(\boldsymbol{y}_s) - \nabla f(\boldsymbol{x}_s)\| \leq \mathcal{L}_s^{(x)}\|\boldsymbol{y}_s - \boldsymbol{x}_s\|, \quad \|\nabla f(\boldsymbol{y}_{s+1}) - \nabla f(\boldsymbol{y}_s)\| \leq \mathcal{L}_s^{(y)}\|\boldsymbol{y}_{s+1} - \boldsymbol{y}_s\|, \tag{67}$$

$$f(\boldsymbol{y}_{s+1}) - f(\boldsymbol{y}_s) - \langle \nabla f(\boldsymbol{y}_s), \boldsymbol{y}_{s+1} - \boldsymbol{y}_s\rangle \leq \frac{\mathcal{L}_s^{(y)}}{2}\|\boldsymbol{y}_{s+1} - \boldsymbol{y}_s\|. \tag{68}$$

In the generalized smooth case, Lemma B.4 does not hold. In contrast, we provide a generalized smooth version of [49, Lemma A.5], which establishes a different relationship between the gradient's norm and the function value gap. Noting that when $q = 1$, Lemma C.3 reduces to [49, Lemma A.5].

Lemma C.3. *Suppose that f is (L_0, L_q)-smooth and Assumption (A1) holds. Then for any $\boldsymbol{x} \in \mathbb{R}^d$,*

$$\|\nabla f(\boldsymbol{x})\| \leq \max\left\{4L_q(f(\boldsymbol{x}) - f^*), [4L_q(f(\boldsymbol{x}) - f^*)]^{\frac{1}{2-q}}, \sqrt{4L_0(f(\boldsymbol{x}) - f^*)}\right\}.$$

C.2 Probabilistic estimations

The probabilistic inequalities in (35) and (36) remain unchanged since they do not rely on any smooth-related conditions. However, we shall rely on a different setting of λ in (36) as follows.

Lemma C.4. *Given $T \geq 1$ and $\delta \in (0, 1)$. Under the same conditions of Lemma B.7, if we set $\lambda = (1 - \beta_1)\sqrt{1 - \beta_2}/(3\eta\mathcal{H})$ where $\mathcal{H}$ is as in* (11)*, then with probability at least $1 - \delta$,*

$$\sum_{s=1}^{t} -\eta_s \left\langle \bar{\boldsymbol{g}}_s, \frac{\boldsymbol{\xi}_s}{\boldsymbol{a}_s} \right\rangle \leq \frac{\mathcal{G}_T(t)}{4\mathcal{H}} \sum_{s=1}^{t} \eta_s \left\| \frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}} \right\|^2 + D_1\mathcal{H}, \quad \forall t \in [T], \tag{69}$$

where D_1 is given in Lemma B.7.

The bounds of the four summations in Lemma B.3 also remain unchanged. However, the upper bound for $\mathcal{F}_i(t)$ should be revised by the following lemma. The detailed proof could be found in Appendix E.

Lemma C.5. *Given $T \geq 1$. Under the conditions and notations of Lemma B.3, if f is (L_0, L_q)-smooth, $\eta = \tilde{C}_0\sqrt{1 - \beta_2}$,* (35) *and* (66) *hold, then the following inequalities hold,*

$$\mathcal{F}_i(t) \leq \mathcal{J}(t), \quad \forall t \in [T], i \in [d], \tag{70}$$

where $\mathcal{J}(t)$ is defined as

$$\mathcal{J}(t) := 1 + \frac{2\mathcal{M}_T^2}{\epsilon^2}\left[\sigma_0^2 t + \sigma_1^2 t\left(\|\bar{\boldsymbol{g}}_1\| + t\tilde{M}_t\right)^p + t\left(\|\bar{\boldsymbol{g}}_1\| + t\tilde{M}_t\right)^2\right], \tag{71}$$

and $\tilde{M}_t := \tilde{C}_0\mathcal{L}_t^{(x)}\sqrt{\frac{d}{1-\beta_1/\beta_2}}$.

It's worth noting that $\mathcal{J}(t)$ is still random relying on the random variable $\mathcal{L}_t^{(x)}$.

C.3 Deterministic estimations

Note that (40) in Appendix B.4 remains unchanged since it's independent from any smooth-related condition. In terms of **A.1**, the only difference is using $\mathcal{H}$ to replace $\mathcal{G}$ in (39) as we choose a different λ in (36), leading to

$$\mathbf{A.1} \leq \left(\frac{\mathcal{G}_T(t)}{4\mathcal{H}} - \frac{3}{4}\right)\sum_{s=1}^{t} \eta_s \left\| \frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}} \right\|^2 + D_1\mathcal{H} + D_2\mathcal{G}_T(t)\sum_{s=1}^{t}\left\| \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s} \right\|^2. \tag{72}$$

We also establish the following proposition which is a generalized smooth version of Proposition B.13.

Proposition C.6. *Given $T \geq 1$. If f is (L_0, L_q)-smooth and* (66) *holds, then*

$$f(\boldsymbol{y}_{t+1}) \leq f(\boldsymbol{x}_1) + \mathbf{A.1} + \mathbf{B.1} + \sum_{s=1}^{t-1} D_6(s)\left\| \frac{\hat{\boldsymbol{m}}_s}{\boldsymbol{b}_s} \right\|^2 + \sum_{s=1}^{t} D_7(s)\left\| \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s} \right\|^2, \quad \forall t \in [T], \tag{73}$$

where $\Sigma_{\max}$ is as in Lemma B.1 and $D_6(s), D_7(s)$ are defined as,[5]

$$D_6(s) = \frac{\mathcal{L}_s^{(y)}\eta^2(1 + 4\Sigma_{\max}^2)}{2(1 - \beta_1)^2}, \quad D_7(s) = \frac{3\mathcal{L}_s^{(y)}\eta^2}{2(1 - \beta_1)^2}.$$

[5] The notations are different from D_6 and D_7 defined in (42).

Proof. The proof follows some same parts in proving Proposition B.14. We start from the descent lemma (68) in Lemma C.2 and sum over $s \in [t]$ to obtain that

$$
\begin{aligned}
f(\boldsymbol{y}_{t+1}) &\le f(\boldsymbol{x}_1) + \sum_{s=1}^{t} \langle \nabla f(\boldsymbol{y}_s), \boldsymbol{y}_{s+1} - \boldsymbol{y}_s \rangle + \sum_{s=1}^{t} \frac{\mathcal{L}_s^{(y)}}{2} \|\boldsymbol{y}_{s+1} - \boldsymbol{y}_s\|^2 \\
&= f(\boldsymbol{x}_1) + \mathbf{A} + \mathbf{B} + \underbrace{\sum_{s=1}^{t} \frac{\mathcal{L}_s^{(y)}}{2} \left\| \eta_s \cdot \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s} - \frac{\beta_1}{1-\beta_1} \left(\frac{\eta_s \boldsymbol{b}_{s-1}}{\eta_{s-1} \boldsymbol{b}_s} - \mathbf{1} \right) \Sigma_s \odot (\boldsymbol{x}_s - \boldsymbol{x}_{s-1}) \right\|^2}_{\textbf{C'}},
\end{aligned}
\tag{74}
$$

where **A** and **B** follow the same definitions in (32). We also follow the decompositions in (33) and (34). We could also rely on the same analysis for the smooth case in (46) but the smooth parameter is replaced by $\mathcal{L}_s^{(x)}$. Hence, we obtain that

$$
\textbf{A.2} \le \sum_{s=1}^{t} \frac{\mathcal{L}_s^{(x)} \eta^2}{2(1-\beta_1)^2} \left\| \frac{\hat{\boldsymbol{m}}_{s-1}}{\boldsymbol{b}_{s-1}} \right\|^2 + \sum_{s=1}^{t} \frac{\mathcal{L}_s^{(x)} \eta^2}{2(1-\beta_1)^2} \left\| \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s} \right\|^2 . \tag{75}
$$

Similarly,

$$
\textbf{B.2} \le \sum_{s=1}^{t} \frac{\Sigma_{\max}^2 \mathcal{L}_s^{(x)} \eta^2}{(1-\beta_1)^2} \left\| \frac{\hat{\boldsymbol{m}}_{s-1}}{\boldsymbol{b}_{s-1}} \right\|^2 , \tag{76}
$$

Noting that **C'** differs from **C** with L replaced by $\mathcal{L}_s^{(y)}$. Hence, relying on a similar analysis in (48), we obtain that

$$
\textbf{C'} \le \sum_{s=1}^{t} \frac{\mathcal{L}_s^{(y)} \eta^2}{(1-\beta_1)^2} \left\| \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s} \right\|^2 + \sum_{s=1}^{t} \frac{\Sigma_{\max}^2 \mathcal{L}_s^{(y)} \eta^2}{(1-\beta_1)^2} \left\| \frac{\hat{\boldsymbol{m}}_{s-1}}{\boldsymbol{b}_{s-1}} \right\|^2 . \tag{77}
$$

Combining (74) with (75), (76) and (77), and noting that $\mathcal{L}_s^{(x)} \le \mathcal{L}_s^{(y)}$ from (64), we thereby obtain the desired result. □

C.4 Bounding gradients

Based on the unchanged parts in Appendix B.3 and Appendix B.4 and the new estimations in (72) and (73), we are now ready to provide the uniform gradients' bound in the following proposition.

Proposition C.7. *Under the same conditions in Theorem 4.1, for any given $\delta \in (0, 1/2)$, it holds that with probability at least $1 - 2\delta$,*

$$
\|\bar{\boldsymbol{g}}_t\| \le H, \quad \mathcal{G}_T(t) \le \mathcal{H}, \quad \mathcal{L}_t^{(x)} \le \mathcal{L}_t^{(y)} \le \mathcal{L}, \quad \forall t \in [T+1], \tag{78}
$$

and

$$
f(\boldsymbol{y}_{t+1}) - f^* \le -\frac{1}{4} \sum_{s=1}^{t} \eta_s \left\| \frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}} \right\|^2 + \hat{H}, \quad \forall t \in [T], \tag{79}
$$

where $H, \mathcal{H}, \mathcal{L}$ are given in (11) *and $\hat{H}$ is given in* (82).

Proof. Based on the two inequalities (35) and (69), we could deduce the final results in (78) and (79). Since (35) and (69) hold with probability at least $1 - 2\delta$, we thereby deduce the desired result holding with probability at least $1 - 2\delta$. To start with, we shall verify that (66) always holds. Recalling η in (10) and F in Lemma C.1,

$$
\eta F = \tilde{C}_0 \sqrt{1-\beta_2} F \le \sqrt{\frac{\beta_2 (1-\beta_1)^2 (1-\beta_2)(1-\beta_1/\beta_2)}{4 L_q^2 d}} \cdot F \le \frac{1}{L_q} .
$$

Hence, we make sure that the distance requirement in (8) always holds according to Lemma C.1. Second, plugging (72) and (40) into the result in (73),

$$
\begin{aligned}
f(\boldsymbol{y}_{t+1}) \leq & f(\boldsymbol{x}_1) + \left(\frac{\mathcal{G}_T(t)}{4\mathcal{H}} - \frac{1}{2}\right)\sum_{s=1}^{t}\eta_s\left\|\frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}}\right\|^2 + D_1\mathcal{H} + D_2\mathcal{G}_T(t)\sum_{s=1}^{t}\left\|\frac{\boldsymbol{g}_s}{\boldsymbol{b}_s}\right\|^2 \\
& + \sum_{s=1}^{t} D_7(s)\left\|\frac{\boldsymbol{g}_s}{\boldsymbol{b}_s}\right\|^2 + (D_3\mathcal{G}_T(t) + D_4)\sum_{s=1}^{t}\left(\left\|\frac{\boldsymbol{m}_{s-1}}{\boldsymbol{b}_s}\right\|^2 + \left\|\frac{\boldsymbol{m}_{s-1}}{\boldsymbol{b}_{s-1}}\right\|^2\right) \\
& + D_5 G_t + \sum_{s=1}^{t-1} D_6(s)\left\|\frac{\hat{\boldsymbol{m}}_s}{\boldsymbol{b}_s}\right\|^2. \qquad (80)
\end{aligned}
$$

We still rely on an induction argument to deduce the result. First, we provide the detail expressions of $\hat{H}, H$ as follows which is determined by hyper-parameters β_1, β_2 and constants $E_0, d, T, \delta, \mathcal{M}_T$,

$$
\begin{aligned}
\hat{H} := & f(\boldsymbol{x}_1) - f^* + \frac{3E_0\mathcal{M}_T}{1-\beta_1}\log\left(\frac{T}{\delta}\right) + \frac{E_0\mathcal{M}_T d}{1-\beta_1}\log\left(\frac{\tilde{\mathcal{J}}(T)}{\beta_2^T}\right) \\
& + \frac{4E_0(\mathcal{M}_T+\epsilon)d}{\beta_2(1-\beta_1)^2(1-\beta_1/\beta_2)}\log\left(\frac{\tilde{\mathcal{J}}(T)}{\beta_2^T}\right) + \frac{2E_0 d}{\sqrt{(1-\beta_1)^3(1-\beta_1/\beta_2)}} \\
& + \frac{3E_0^2 d}{2(1-\beta_1)^2}\log\left(\frac{\tilde{\mathcal{J}}(T)}{\beta_2^T}\right) + \frac{5E_0^2 d}{2\beta_2(1-\beta_1)^2(1-\beta_1/\beta_2)}\log\left(\frac{\tilde{\mathcal{J}}(T)}{\beta_2^T}\right), \qquad (81)
\end{aligned}
$$

$$
H := L_0/L_q + \left(4L_q\hat{H}\right)^q + \left(4L_q\hat{H}\right)^{\frac{q}{2-q}} + \left(4L_0\hat{H}\right)^{\frac{q}{2}} + 4L_q\hat{H} + \left(4L_q\hat{H}\right)^{\frac{1}{2-q}} + \sqrt{4L_0\hat{H}}. \qquad (82)
$$

where $E_0 > 0$ is a constant and $\tilde{\mathcal{J}}(T)$ is a polynomial of T given as

$$
\tilde{\mathcal{J}}(T) := 1 + \frac{2\mathcal{M}_T^2}{\epsilon^2}\left[\sigma_0^2 T + \sigma_1^2 T\left(\|\bar{\boldsymbol{g}}_1\| + T\tilde{M}\right)^p + T\left(\|\bar{\boldsymbol{g}}_1\| + T\tilde{M}\right)^2\right], \qquad (83)
$$

and $\tilde{M} := E_0\sqrt{\frac{d}{1-\beta_1/\beta_2}}$. The induction then begins by noting that from Lemma C.3 and H in (82),

$$
G_1 = \|\bar{\boldsymbol{g}}_1\| \leq 4L_q(f(\boldsymbol{x}_1) - f^*) + (4L_q(f(\boldsymbol{x}_1) - f^*))^{\frac{1}{2-q}} + \sqrt{4L_0(f(\boldsymbol{x}_1) - f^*)} \leq H.
$$

Suppose that for some $t \in [T]$,

$$
G_s \leq H, \quad \forall s \in [t]. \qquad (84)
$$

Consequently, recalling $\mathcal{G}_T(s)$ in (19), $\mathcal{L}_s^{(x)}, \mathcal{L}_s^{(y)}$ in (64) and $\mathcal{H}, \mathcal{L}$ in (11),

$$
\mathcal{G}_T(s) \leq \mathcal{H}, \quad \mathcal{L}_s^{(x)} \leq \mathcal{L}_s^{(y)} \leq \mathcal{L}, \quad \forall s \in [t]. \qquad (85)
$$

We thus apply (85) to (80),

$$
\begin{aligned}
f(\boldsymbol{y}_{t+1}) \leq & f(\boldsymbol{x}_1) - \frac{1}{4}\sum_{s=1}^{t}\eta_s\left\|\frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}}\right\|^2 + D_1\mathcal{H} + D_2\mathcal{H}\sum_{s=1}^{t}\left\|\frac{\boldsymbol{g}_s}{\boldsymbol{b}_s}\right\|^2 + \sum_{s=1}^{t} D_7(s)\left\|\frac{\boldsymbol{g}_s}{\boldsymbol{b}_s}\right\|^2 \\
& + (D_3\mathcal{H} + D_4)\sum_{s=1}^{t}\left(\left\|\frac{\boldsymbol{m}_{s-1}}{\boldsymbol{b}_s}\right\|^2 + \left\|\frac{\boldsymbol{m}_{s-1}}{\boldsymbol{b}_{s-1}}\right\|^2\right) + D_5 H + \sum_{s=1}^{t-1} D_6(s)\left\|\frac{\hat{\boldsymbol{m}}_s}{\boldsymbol{b}_s}\right\|^2. \qquad (86)
\end{aligned}
$$

Further recalling the setting of $\tilde{C}_0$ in (10), with a simple calculation it holds that,

$$
\tilde{C}_0 H \leq E_0, \quad \tilde{C}_0\mathcal{H} \leq E_0, \quad \tilde{C}_0\mathcal{L} \leq E_0, \quad \tilde{C}_0^2\mathcal{L} \leq E_0^2, \quad \tilde{C}_0\epsilon_0 \leq E_0\epsilon_0. \qquad (87)
$$

Therefore, combining with (85), (87) and $\tilde{M}_t$ in (71), we could use the deterministic polynomial $\tilde{\mathcal{J}}(t)$ to further control $\mathcal{J}(t)$ in (71),

$$
\begin{aligned}
& \tilde{M}_t \leq \tilde{C}_0\mathcal{L}\sqrt{\frac{d}{1-\beta_1/\beta_2}} \leq E_0\sqrt{\frac{d}{1-\beta_1/\beta_2}} = \tilde{M}, \quad \mathcal{J}(t) \leq \tilde{\mathcal{J}}(t) \leq \tilde{\mathcal{J}}(T), \\
& \log\left(\frac{\mathcal{F}_i(t)}{\beta_2^t}\right) \leq \log\left(\frac{\mathcal{J}(t)}{\beta_2^t}\right) \leq \log\left(\frac{\tilde{\mathcal{J}}(T)}{\beta_2^T}\right), \quad \forall t \leq T, i \in [d].
\end{aligned}
$$

Then, we could use $\tilde{\mathcal{J}}(T)$ to control the four summations in Lemma B.3 which emerge in (86). In addition, we rely on $\eta = \tilde{C}_0\sqrt{1-\beta_2}$ and the induction assumptions of (84) and (85) to further upper bound the RHS of (86), leading to

$$\begin{aligned} f(\boldsymbol{y}_{t+1}) - f^* \leq & f(\boldsymbol{x}_1) - f^* - \frac{1}{4}\sum_{s=1}^{t} \eta_s \left\| \frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}} \right\|^2 + \frac{3\tilde{C}_0\mathcal{H}}{1-\beta_1}\log\left(\frac{T}{\delta}\right) + \frac{\tilde{C}_0\mathcal{H}d}{1-\beta_1}\log\left(\frac{\tilde{\mathcal{J}}(T)}{\beta_2^T}\right) \\ & + \frac{4\tilde{C}_0(\mathcal{H}+\epsilon_0)d}{\beta_2(1-\beta_1)^2(1-\beta_1/\beta_2)}\log\left(\frac{\tilde{\mathcal{J}}(T)}{\beta_2^T}\right) + \frac{2\tilde{C}_0 H d}{\sqrt{(1-\beta_1)^3(1-\beta_1/\beta_2)}} \\ & + \frac{3\tilde{C}_0^2\mathcal{L}d}{2(1-\beta_1)^2}\log\left(\frac{\tilde{\mathcal{J}}(T)}{\beta_2^T}\right) + \frac{5\tilde{C}_0^2\mathcal{L}d}{2\beta_2(1-\beta_1)^2(1-\beta_1/\beta_2)}\log\left(\frac{\tilde{\mathcal{J}}(T)}{\beta_2^T}\right). \end{aligned} \tag{88}$$

Then combining with (87) and the definition of $\hat{H}$ in (81), we obtain that

$$\Delta_{t+1} := f(\boldsymbol{y}_{t+1}) - f^* \leq -\frac{1}{4}\sum_{s=1}^{t}\eta_s \left\| \frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}} \right\|^2 + \hat{H} \leq \hat{H}. \tag{89}$$

Then, further using Lemma C.2, Lemma C.3 and H in (82),

$$\begin{aligned} \|\bar{\boldsymbol{g}}_{t+1}\| &\leq L_0/L_q + \|\nabla f(\boldsymbol{y}_{t+1})\|^q + \|\nabla f(\boldsymbol{y}_{t+1})\| \\ &\leq L_0/L_q + (4L_q\Delta_{t+1})^q + (4L_q\Delta_{t+1})^{\frac{q}{2-q}} \\ &+ (4L_0\Delta_{t+1})^{\frac{q}{2}} + 4L_q\Delta_{t+1} + (4L_q\Delta_{t+1})^{\frac{1}{2-q}} + \sqrt{4L_0\Delta_{t+1}} \leq H. \end{aligned}$$

We then deduce that $G_{t+1} = \max\{G_t, \|\bar{\boldsymbol{g}}_{t+1}\|\} \leq H$. The induction is then complete and we obtain the desired result in (78). Finally, as an intermediate result of the proof, we obtain that (79) holds as well.

□

C.5 Proof of the main result

Proof of Theorem 4.1. The proof for the final convergence rate follows a similar idea and some same estimations in the proof of Theorem 3.1. Setting $t = T$ in (79), it holds that with probability at least $1 - 2\delta$,

$$\frac{1}{4}\sum_{s=1}^{t}\frac{\eta_s}{\|\boldsymbol{a}_s\|_\infty}\cdot\|\bar{\boldsymbol{g}}_s\|^2 \leq \frac{1}{4}\sum_{s=1}^{t}\eta_s\left\|\frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}}\right\|^2 \leq \hat{H}. \tag{90}$$

Then in what follows, we would assume that (78) and (90) always hold. Relying on the two inequalities, we thereby deduce the final convergence result. Furthermore, since (78) and (90) hold with probability at least $1-2\delta$, the final convergence result also holds with probability at least $1-2\delta$. Using (78) and following the same analysis in (60),

$$\|\boldsymbol{a}_s\|_\infty \leq \max_{i\in[d]}\sqrt{(1-\beta_2)\left(\sum_{j=1}^{s-1}\beta_2^{s-j}g_{j,i}^2 + (\mathcal{G}_T(j))^2\right)} + \epsilon_s \leq (\mathcal{H}+\epsilon)\sqrt{1-\beta_2^s}, \quad \forall s \in [T].$$

Combining with the parameter setting in (10),

$$\frac{\eta_s}{\|\boldsymbol{a}_s\|_\infty} \geq \frac{\eta\sqrt{1-\beta_2^s}}{(1-\beta_1^s)\|\boldsymbol{a}_s\|_\infty} \geq \frac{\tilde{C}_0\sqrt{1-\beta_2}}{\mathcal{H}+\epsilon_0\sqrt{1-\beta_2}}.$$

We then combine with (90) and $\mathcal{H}$ in (11) to obtain that with probability at least $1-2\delta$,

$$\frac{1}{T}\sum_{s=1}^{T}\|\bar{\boldsymbol{g}}_s\|^2 \leq \frac{4\hat{H}}{T\tilde{C}_0}\left(\frac{\sqrt{2(\sigma_0^2+\sigma_1^2H^p+H^2)}}{\sqrt{1-\beta_2}} + \epsilon_0\right)\sqrt{\log\left(\frac{\mathrm{e}T}{\delta}\right)}.$$

Finally, following the same deduction in (62) and (63), we could derive that $\hat{H} \sim \mathcal{O}\left(\log^2\left(\frac{T}{\epsilon_0\delta}\right)\right)$ from (81) and the desired results in Theorem 4.1. □

D Omitted proof in Appendix B

D.1 Omitted proof in Appendix B.1

Proof of Lemma B.1. We fix arbitrary $i \in [d]$ and have the following two cases. When $\frac{\eta_s b_{s-1,i}}{\eta_{s-1} b_{s,i}} < 1$, we have

$$\left| \frac{\eta_s b_{s-1,i}}{\eta_{s-1} b_{s,i}} - 1 \right| = 1 - \frac{\eta_s b_{s-1,i}}{\eta_{s-1} b_{s,i}} < 1.$$

When $\frac{\eta_s b_{s-1,i}}{\eta_{s-1} b_{s,i}} \geq 1$, let $r = \beta_2^{s-1}$. Since $0 < 1 - \beta_1^{s-1} < 1 - \beta_1^s, \forall s \geq 2$, then we have

$$\frac{\eta_s}{\eta_{s-1}} = \sqrt{\frac{1-\beta_2^s}{1-\beta_2^{s-1}} \cdot \frac{1-\beta_1^{s-1}}{1-\beta_1^s}} \leq \sqrt{1 + \frac{\beta_2^{s-1}(1-\beta_2)}{1-\beta_2^{s-1}}} = \sqrt{1 + (1-\beta_2) \cdot \frac{r}{1-r}}.$$

Since $h(r) = r/(1-r)$ is increasing as r grows and r takes the maximum value when $s = 2$. Hence, it holds that

$$\frac{\eta_s}{\eta_{s-1}} \leq \sqrt{1 + (1-\beta_2) \cdot \frac{\beta_2}{1-\beta_2}} = \sqrt{1+\beta_2}. \tag{91}$$

Then, since $\epsilon_{s-1} \leq \epsilon_s$, we further have

$$\frac{b_{s-1,i}}{b_{s,i}} = \frac{\epsilon_{s-1} + \sqrt{v_{s-1,i}}}{\epsilon_s + \sqrt{\beta_2 v_{s-1,i} + (1-\beta_2) g_{s,i}^2}} \leq \frac{\epsilon_s + \sqrt{v_{s-1,i}}}{\epsilon_s + \sqrt{\beta_2 v_{s-1,i}}} \leq \frac{1}{\sqrt{\beta_2}}. \tag{92}$$

Combining with (91) and (92), we have

$$\left| \frac{\eta_s b_{s-1,i}}{\eta_{s-1} b_{s,i}} - 1 \right| = \frac{\eta_s b_{s-1,i}}{\eta_{s-1} b_{s,i}} - 1 \leq \sqrt{\frac{1+\beta_2}{\beta_2}} - 1.$$

Combining the two cases and noting that the bound holds for any $i \in [d]$, we then obtain the desired result. □

Proof of Lemma B.2. Denoting $\tilde{M} = \sum_{j=1}^{s-1} \beta_1^{s-1-j}$ and applying (28) with $\hat{M}$ and α_j replaced by $\tilde{M}$ and $g_{j,i}$ respectively,

$$\left(\sum_{j=1}^{s-1} \beta_1^{s-1-j} g_{j,i} \right)^2 \leq \tilde{M} \cdot \sum_{j=1}^{s-1} \beta_1^{s-1-j} g_{j,i}^2. \tag{93}$$

Hence, combining with the definition of $b_{s,i}$ in (16), we further have for any $i \in [d]$ and $s \geq 2$,

$$\begin{aligned}
\left| \frac{m_{s-1,i}}{b_{s-1,i}} \right| \leq \left| \frac{m_{s-1,i}}{\sqrt{v_{s-1,i}}} \right| &= \sqrt{\frac{(1-\beta_1)^2 \left(\sum_{j=1}^{s-1} \beta_1^{s-1-j} g_{j,i} \right)^2}{(1-\beta_2) \sum_{j=1}^{s-1} \beta_2^{s-1-j} g_{j,i}^2}} \\
&\leq \frac{1-\beta_1}{\sqrt{1-\beta_2}} \sqrt{\tilde{M} \cdot \frac{\sum_{j=1}^{s-1} \beta_1^{s-1-j} g_{j,i}^2}{\sum_{j=1}^{s-1} \beta_2^{s-1-j} g_{j,i}^2}} \leq \frac{1-\beta_1}{\sqrt{1-\beta_2}} \sqrt{\tilde{M} \cdot \sum_{j=1}^{s-1} \left(\frac{\beta_1}{\beta_2} \right)^{s-1-j}} \\
&= \frac{1-\beta_1}{\sqrt{1-\beta_2}} \sqrt{\frac{1-\beta_1^{s-1}}{1-\beta_1} \cdot \frac{1-(\beta_1/\beta_2)^{s-1}}{1-\beta_1/\beta_2}} \leq \sqrt{\frac{(1-\beta_1)(1-\beta_1^{s-1})}{(1-\beta_2)(1-\beta_1/\beta_2)}},
\end{aligned}$$

where the last inequality applies $\beta_1 < \beta_2$. We thus prove the first result. To prove the second result, from the smoothness of f,

$$\|\bar{\boldsymbol{g}}_s\| \leq \|\bar{\boldsymbol{g}}_{s-1}\| + \|\bar{\boldsymbol{g}}_s - \bar{\boldsymbol{g}}_{s-1}\| \leq \|\bar{\boldsymbol{g}}_{s-1}\| + L\|\boldsymbol{x}_s - \boldsymbol{x}_{s-1}\|. \tag{94}$$

Combining with (30) and $\eta = C_0\sqrt{1-\beta_2}$,

$$\|\boldsymbol{x}_s - \boldsymbol{x}_{s-1}\|_\infty \le \eta_{s-1}\left\|\frac{\boldsymbol{m}_{s-1}}{\boldsymbol{b}_{s-1}}\right\|_\infty \le \eta\sqrt{\frac{1}{(1-\beta_2)(1-\beta_1/\beta_2)}} = C_0\sqrt{\frac{1}{1-\beta_1/\beta_2}}. \tag{95}$$

Using $\|\boldsymbol{x}_s - \boldsymbol{x}_{s-1}\| \le \sqrt{d}\|\boldsymbol{x}_s - \boldsymbol{x}_{s-1}\|_\infty$ and (94),

$$\|\bar{\boldsymbol{g}}_s\| \le \|\bar{\boldsymbol{g}}_{s-1}\| + LC_0\sqrt{\frac{d}{1-\beta_1/\beta_2}} \le \|\bar{\boldsymbol{g}}_1\| + LC_0 s\sqrt{\frac{d}{1-\beta_1/\beta_2}}.$$

□

Proof of Lemma B.3. Recalling the updated rule and the definition of $b_{s,i}$ in (16), using $\epsilon_s^2 = \epsilon^2(1-\beta_2^s) \ge \epsilon^2(1-\beta_2)$,

$$b_{s,i}^2 \ge v_{s,i}^2 + \epsilon_s^2 \ge (1-\beta_2)\left(\sum_{j=1}^{s}\beta_2^{s-j}g_{j,i}^2 + \epsilon^2\right), \quad \text{and} \quad m_{s,i} = (1-\beta_1)\sum_{j=1}^{s}\beta_1^{s-j}g_{j,i}. \tag{96}$$

Proof for the first summation Using (96), for any $i \in [d]$,

$$\sum_{s=1}^{t}\frac{g_{s,i}^2}{b_{s,i}^2} \le \frac{1}{1-\beta_2}\sum_{s=1}^{t}\frac{g_{s,i}^2}{\epsilon^2 + \sum_{j=1}^{s}\beta_2^{s-j}g_{j,i}^2}.$$

Applying Lemma A.1 and recalling the definition of $\mathcal{F}_i(t)$,

$$\sum_{s=1}^{t}\frac{g_{s,i}^2}{b_{s,i}^2} \le \frac{1}{1-\beta_2}\left[\log\left(1 + \frac{1}{\epsilon^2}\sum_{s=1}^{t}\beta_2^{t-s}g_{s,i}^2\right) - t\log\beta_2\right] \le \frac{1}{1-\beta_2}\log\left(\frac{\mathcal{F}_i(t)}{\beta_2^t}\right).$$

Summing over $i \in [d]$, we obtain the first desired result.

Proof for the second summation Following from (96),

$$\sum_{s=1}^{t}\frac{m_{s,i}^2}{b_{s,i}^2} \le \frac{(1-\beta_1)^2}{1-\beta_2}\cdot\sum_{s=1}^{t}\frac{\left(\sum_{j=1}^{s}\beta_1^{s-j}g_{j,i}\right)^2}{\epsilon^2 + \sum_{j=1}^{s}\beta_2^{s-j}g_{j,i}^2}.$$

Applying Lemma A.2 and $\beta_2 \le 1$,

$$\begin{aligned}\sum_{s=1}^{t}\frac{m_{s,i}^2}{b_{s,i}^2} &\le \frac{(1-\beta_1)^2}{1-\beta_2}\cdot\frac{1}{(1-\beta_1)(1-\beta_1/\beta_2)}\left[\log\left(1 + \frac{1}{\epsilon^2}\sum_{s=1}^{t}\beta_2^{t-s}g_{s,i}^2\right) - t\log\beta_2\right]\\ &= \frac{1-\beta_1}{(1-\beta_2)(1-\beta_1/\beta_2)}\log\left(\frac{\mathcal{F}_i(t)}{\beta_2^t}\right).\end{aligned}$$

Summing over $i \in [d]$, we obtain the second desired result.

Proof for the third summation Following from (96),

$$\begin{aligned}\sum_{s=1}^{t}\frac{m_{s,i}^2}{b_{s+1,i}^2} &\le \sum_{s=1}^{t}\frac{\left[(1-\beta_1)\sum_{j=1}^{s}\beta_1^{s-j}g_{j,i}\right]^2}{\epsilon^2(1-\beta_2) + (1-\beta_2)\sum_{j=1}^{s+1}\beta_2^{s+1-j}g_{j,i}^2}\\ &\le \sum_{s=1}^{t}\frac{(1-\beta_1)^2\left(\sum_{j=1}^{s}\beta_1^{s-j}g_{j,i}\right)^2}{\epsilon^2(1-\beta_2) + (1-\beta_2)\beta_2\sum_{j=1}^{s}\beta_2^{s-j}g_{j,i}^2}.\end{aligned}$$

Applying Lemma A.2, and using $\beta_2 \le 1$,

$$\begin{aligned}
\sum_{s=1}^{t} \frac{m_{s,i}^2}{b_{s+1,i}^2} &\le \frac{(1-\beta_1)^2}{(1-\beta_2)\beta_2} \cdot \sum_{s=1}^{t} \frac{\left(\sum_{j=1}^{s} \beta_1^{s-j} g_{j,i}\right)^2}{\frac{\epsilon^2}{\beta_2} + \sum_{j=1}^{s} \beta_2^{s-j} g_{j,i}^2} \\
&\le \frac{(1-\beta_1)^2}{(1-\beta_2)\beta_2} \cdot \frac{1}{(1-\beta_1)(1-\beta_1/\beta_2)} \left[\log\left(1 + \frac{\beta_2}{\epsilon^2} \sum_{s=1}^{t} \beta_2^{t-s} g_{s,i}^2 \right) - t \log \beta_2 \right] \\
&\le \frac{1-\beta_1}{\beta_2(1-\beta_2)(1-\beta_1/\beta_2)} \log\left(\frac{\mathcal{F}_i(t)}{\beta_2^t}\right).
\end{aligned}$$

Summing over $i \in [d]$, we obtain the third desired result.

Proof for the fourth summation Following the definition of $\hat{m}_{s,i}$ from (29), and combining with (96),

$$\sum_{s=1}^{t} \frac{\hat{m}_{s,i}^2}{b_{s,i}^2} \le \frac{(1-\beta_1)^2}{1-\beta_2} \cdot \sum_{s=1}^{t} \frac{\left(\frac{1}{1-\beta_1^s} \sum_{j=1}^{s} \beta_1^{s-j} g_{j,i}\right)^2}{\epsilon^2 + \sum_{j=1}^{s} \beta_2^{s-j} g_{j,i}^2}.$$

Applying Lemma A.2 and using $\beta_2 \le 1$,

$$\begin{aligned}
\sum_{s=1}^{t} \frac{\hat{m}_{s,i}^2}{b_{s,i}^2} &\le \frac{(1-\beta_1)^2}{1-\beta_2} \cdot \frac{1}{(1-\beta_1)^2(1-\beta_1/\beta_2)} \left[\log\left(1 + \frac{1}{\epsilon^2} \sum_{s=1}^{t} \beta_2^{t-s} g_{s,i}^2 \right) - t \log \beta_2 \right] \\
&\le \frac{1}{(1-\beta_2)(1-\beta_1/\beta_2)} \log\left(\frac{\mathcal{F}_i(t)}{\beta_2^t}\right).
\end{aligned}$$

Summing over $i \in [d]$, we obtain the fourth desired result. □

Proof of Lemma B.4. Let $\hat{\boldsymbol{x}} = \boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x})$. Then using the descent lemma of smoothness,

$$f(\hat{\boldsymbol{x}}) \le f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \hat{\boldsymbol{x}} - \boldsymbol{x} \rangle + \frac{L}{2}\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|^2 \le f(\boldsymbol{x}) - \frac{1}{2L}\|\nabla f(\boldsymbol{x})\|^2.$$

Re-arranging the order, and noting that $f(\hat{\boldsymbol{x}}) \ge f^*$,

$$\|\nabla f(\boldsymbol{x})\|^2 \le 2L(f(\boldsymbol{x}) - f(\hat{\boldsymbol{x}})) \le 2L(f(\boldsymbol{x}) - f^*).$$

□

Proof of Lemma B.5. Applying the norm inequality and the smoothness of f,

$$\|\nabla f(\boldsymbol{x}_s)\| \le \|\nabla f(\boldsymbol{y}_s)\| + \|\nabla f(\boldsymbol{x}_s) - \nabla f(\boldsymbol{y}_s)\| \le \|\nabla f(\boldsymbol{y}_s)\| + L\|\boldsymbol{y}_s - \boldsymbol{x}_s\|.$$

Combining with the definition of $\boldsymbol{y}_s$ in (17) and (95), and using $\beta_1 \in [0,1)$, we obtain the desired result that

$$\|\nabla f(\boldsymbol{x}_s)\| \le \|\nabla f(\boldsymbol{y}_s)\| + \frac{L\beta_1}{1-\beta_1}\|\boldsymbol{x}_s - \boldsymbol{x}_{s-1}\| \le \|\nabla f(\boldsymbol{y}_s)\| + \frac{LC_0\sqrt{d}}{(1-\beta_1)\sqrt{1-\beta_1/\beta_2}}.$$

□

D.2 Omitted proof in Appendix B.3

Proof of Lemma B.6. Let us denote $\gamma_s = \frac{\|\boldsymbol{\xi}_s\|^2}{\sigma_0^2 + \sigma_1^2\|\bar{\boldsymbol{g}}_s\|^p}, \forall s \in [T]$. Then from Assumption (A3), we first have $\mathbb{E}_{\boldsymbol{z}_s}[\exp(\gamma_s)] \le \exp(1)$. Taking full expectation,

$$\mathbb{E}[\exp(\gamma_s)] \le \exp(1).$$

By Markov's inequality, for any $A \in \mathbb{R}$,

$$\mathbb{P}\left(\max_{s\in[T]} \gamma_s \geq A\right) = \mathbb{P}\left(\exp\left(\max_{s\in[T]} \gamma_s\right) \geq \exp(A)\right) \leq \exp(-A)\mathbb{E}\left[\exp\left(\max_{s\in[T]} \gamma_s\right)\right]$$
$$\leq \exp(-A)\mathbb{E}\left[\sum_{s=1}^{T} \exp\left(\gamma_s\right)\right] \leq \exp(-A) T \exp(1),$$

which leads to that with probability at least $1-\delta$,

$$\|\boldsymbol{\xi}_s\|^2 \leq \log\left(\frac{\mathrm{e}T}{\delta}\right)\left(\sigma_0^2 + \sigma_1^2\|\bar{\boldsymbol{g}}_s\|^p\right), \quad \forall s \in [T].$$

□

Proof of Lemma B.7. Recalling the definitions of $\boldsymbol{a}_s$ in (23) and $\boldsymbol{\epsilon}_s$ in Algorithm 1, we have for any $s \in [T], i \in [d]$,

$$\frac{1}{a_{s,i}} \leq \frac{1}{\mathcal{G}_T(s)\sqrt{1-\beta_2} + \epsilon\sqrt{1-\beta_2^s}} \leq \frac{1}{(\mathcal{G}_T(s)+\epsilon)\sqrt{1-\beta_2}}$$
$$\leq \frac{1}{\mathcal{G}_T(s)\sqrt{1-\beta_2}} \leq \frac{1}{\sqrt{\sigma_0^2+\sigma_1^2\|\bar{\boldsymbol{g}}_s\|^p}\sqrt{1-\beta_2}}. \tag{97}$$

Then given any $i \in [d]$, we set

$$X_s = -\eta_s\left\langle \bar{\boldsymbol{g}}_s, \frac{\boldsymbol{\xi}_s}{\boldsymbol{a}_s}\right\rangle, \omega_s = \eta_s\left\|\frac{\bar{\boldsymbol{g}}_s}{\boldsymbol{a}_s}\right\|\sqrt{\sigma_0^2+\sigma_1^2\|\bar{\boldsymbol{g}}_s\|^p}, \quad \forall s \in [T].$$

Noting that $\bar{\boldsymbol{g}}_s, \boldsymbol{a}_s$ and η_s are random variables dependent by $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{s-1}$ and $\boldsymbol{\xi}_s$ is only dependent on $\boldsymbol{z}_s$. We then verify that X_s is a martingale difference sequence since

$$\mathbb{E}\left[X_s \mid \boldsymbol{z}_1, \cdots, \boldsymbol{z}_{s-1}\right] = \mathbb{E}_{\boldsymbol{z}_s}\left[-\eta_s\left\langle \bar{\boldsymbol{g}}_s, \frac{\boldsymbol{\xi}_s}{\boldsymbol{a}_s}\right\rangle\right] = -\eta_s\left\langle \bar{\boldsymbol{g}}_s, \frac{\mathbb{E}_{\boldsymbol{z}_s}[\boldsymbol{\xi}_s]}{\boldsymbol{a}_s}\right\rangle = 0.$$

Noting that ω_s is a random variable only dependent by $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{s-1}$ and applying Assumption (A3) and Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left[\exp\left(\frac{X_s^2}{\omega_s^2}\right) \mid \boldsymbol{z}_1, \cdots, \boldsymbol{z}_{s-1}\right] \leq \mathbb{E}\left[\exp\left(\frac{\boldsymbol{\xi}_s^2}{\sigma_0^2+\sigma_1^2\|\bar{\boldsymbol{g}}_s\|^p}\right) \mid \boldsymbol{z}_1, \cdots, \boldsymbol{z}_{s-1}\right]$$
$$\leq \mathbb{E}_{\boldsymbol{z}_s}\left[\exp\left(\frac{\|\boldsymbol{\xi}_s\|^2}{\sigma_0^2+\sigma_1^2\|\bar{\boldsymbol{g}}_s\|^p}\right)\right] \leq \exp(1), \quad \forall s \in [T].$$

Applying Lemma A.3 and (97), we have that for any $\lambda > 0$, with probability at least $1-\delta$,

$$\sum_{s=1}^{t} X_s \leq \frac{3\lambda}{4}\sum_{s=1}^{t}\omega_s^2 + \frac{1}{\lambda}\log\left(\frac{1}{\delta}\right)$$
$$\leq \frac{3\lambda}{4\sqrt{1-\beta_2}}\sum_{s=1}^{t}\eta_s^2\left\|\frac{\bar{\boldsymbol{g}}_s}{\boldsymbol{a}_s}\right\|^2(\sigma_0^2+\sigma_1^2\|\bar{\boldsymbol{g}}_s\|^p) + \frac{1}{\lambda}\log\left(\frac{1}{\delta}\right)$$
$$\leq \frac{3\lambda}{4\sqrt{1-\beta_2}}\sum_{s=1}^{t}\eta_s^2\left\|\frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}}\right\|^2\sqrt{\sigma_0^2+\sigma_1^2\|\bar{\boldsymbol{g}}_s\|^p} + \frac{1}{\lambda}\log\left(\frac{1}{\delta}\right). \tag{98}$$

Note that for any $t \in [T]$, (98) holds with probability at least $1-\delta$. Then for any fixed $\lambda > 0$, we could re-scale δ to obtain that with probability at least $1-\delta$, for all $t \in [T]$,

$$\sum_{s=1}^{t} X_s \leq \frac{3\lambda}{4\sqrt{1-\beta_2}}\sum_{s=1}^{t}\eta_s^2\left\|\frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}}\right\|^2\sqrt{\sigma_0^2+\sigma_1^2\|\bar{\boldsymbol{g}}_s\|^p} + \frac{1}{\lambda}\log\left(\frac{T}{\delta}\right).$$

Using $\sqrt{\sigma_0^2+\sigma_1^2\|\bar{\boldsymbol{g}}_s\|^p} \leq \mathcal{G}_T(t), s \leq t$ from (19), together with (30), we have that with probability at least $1-\delta$,

$$-\sum_{s=1}^{t}\eta_s\left\langle \bar{\boldsymbol{g}}_s, \frac{\boldsymbol{\xi}_s}{\boldsymbol{a}_s}\right\rangle \leq \frac{3\lambda\eta\mathcal{G}_T(t)}{4(1-\beta_1)\sqrt{1-\beta_2}}\sum_{s=1}^{t}\eta_s\left\|\frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}}\right\|^2 + \frac{1}{\lambda}\log\left(\frac{T}{\delta}\right), \quad \forall t \in [T].$$

Finally setting $\lambda = (1-\beta_1)\sqrt{1-\beta_2}/\left(3\eta\mathcal{G}_T\right)$, we then have the desired result in (37). □

D.3 Omitted proof of Appendix B.4

Proof of Lemma B.8. First directly applying (35) and G_s in (19), for any $j \in [s]$,

$$\|\boldsymbol{\xi}_j\| \le \mathcal{M}_T\sqrt{\sigma_0^2 + \sigma_1^2\|\bar{\boldsymbol{g}}_j\|^p} \le \mathcal{M}_T\sqrt{\sigma_0^2 + \sigma_1^2 G_j^p} \le \mathcal{M}_T\sqrt{\sigma_0^2 + \sigma_1^2 G_s^p} \le \mathcal{G}_T(s).$$

Applying the basic inequality, (35) and $\mathcal{M}_T \ge 1$, for any $j \in [s]$,

$$\|\boldsymbol{g}_j\|^2 \le 2\|\bar{\boldsymbol{g}}_j\|^2 + 2\|\boldsymbol{\xi}_j\|^2 \le 2\mathcal{M}_T^2\left(\sigma_0^2 + \sigma_1^2\|\bar{\boldsymbol{g}}_j\|^p + \|\bar{\boldsymbol{g}}_j\|^2\right) \le (\mathcal{G}_T(s))^2.$$

Finally, we would use an induction argument to prove the last result. Given any $i \in [d]$, noting that $v_{1,i} = (1-\beta_2)g_{1,i}^2 \le (\mathcal{G}_T(s))^2$. Suppose that for some $s' \in [s]$, $v_{j,i} \le (\mathcal{G}_T(s))^2, \forall j \in [s']$,

$$v_{s'+1,i} = \beta_2 v_{s',i} + (1-\beta_2)g_{s',i}^2 \le \beta_2(\mathcal{G}_T(s))^2 + (1-\beta_2)(\mathcal{G}_T(s))^2 \le (\mathcal{G}_T(s))^2.$$

We then obtain that $v_{j,i} \le (\mathcal{G}_T(s))^2, \forall j \in [s]$. Noting that the above inequality holds for all $i \in [d]$, we therefore obtain the desired result. □

Proof of Lemma B.9. Recalling the definition of $b_{s,i}$ in (16) and letting $a_{s,i} = \sqrt{\tilde{v}_{s,i}} + \epsilon_s$ in (23),

$$\begin{aligned}\left|\frac{1}{a_{s,i}} - \frac{1}{b_{s,i}}\right| &= \frac{\left|\sqrt{v_{s,i}} - \sqrt{\tilde{v}_{s,i}}\right|}{a_{s,i}b_{s,i}} = \frac{1-\beta_2}{a_{s,i}b_{s,i}}\frac{\left|g_{s,i}^2 - (\mathcal{G}_T(s))^2\right|}{\sqrt{v_{s,i}} + \sqrt{\tilde{v}_{s,i}}} \\ &\le \frac{1-\beta_2}{a_{s,i}b_{s,i}} \cdot \frac{(\mathcal{G}_T(s))^2}{\sqrt{v_{s,i}} + \sqrt{\beta_2 v_{s-1,i} + (1-\beta_2)(\mathcal{G}_T(s))^2}} \le \frac{\mathcal{G}_T(s)\sqrt{1-\beta_2}}{a_{s,i}b_{s,i}},\end{aligned}$$

where we apply $g_{s,i}^2 \le \|\boldsymbol{g}_s\|^2 \le (\mathcal{G}_T(s))^2$ from Lemma B.8 in the first inequality since (35) holds. The second result also follows from the same analysis. We first combine with $\epsilon_s = \epsilon\sqrt{1-\beta_2^s}$ to obtain that

$$|\epsilon_s - \epsilon_{s-1}| \le \epsilon\left(\sqrt{1-\beta_2^s} - \sqrt{1-\beta_2^{s-1}}\right) \le \epsilon\sqrt{\beta_2^{s-1}(1-\beta_2)} \le \epsilon\sqrt{1-\beta_2}, \tag{99}$$

where we apply $\sqrt{a} - \sqrt{b} \le \sqrt{a-b}, \forall 0 \le b \le a$. Applying the definition of $b_{s-1,i}$ and $a_{s,i}$,

$$\begin{aligned}\left|\frac{1}{b_{s-1,i}} - \frac{1}{a_{s,i}}\right| &= \frac{\left|\sqrt{\tilde{v}_{s,i}} - \sqrt{v_{s-1,i}} + (\epsilon_s - \epsilon_{s-1})\right|}{b_{s-1,i}a_{s,i}} \\ &\le \frac{1}{b_{s-1,i}a_{s,i}}\frac{(1-\beta_2)\left|(\mathcal{G}_T(s))^2 - v_{s-1,i}\right|}{\sqrt{\tilde{v}_{s,i}} + \sqrt{v_{s-1,i}}} + \frac{|\epsilon_s - \epsilon_{s-1}|}{b_{s-1,i}a_{s,i}} \\ &\le \frac{1}{b_{s-1,i}a_{s,i}} \cdot \frac{(1-\beta_2)(\mathcal{G}_T(s))^2}{\sqrt{\tilde{v}_{s,i}} + \sqrt{v_{s-1,i}}} + \frac{\epsilon\sqrt{1-\beta_2}}{b_{s-1,i}a_{s,i}} \le \frac{(\mathcal{G}_T(s) + \epsilon)\sqrt{1-\beta_2}}{b_{s-1,i}a_{s,i}}.\end{aligned}$$

where the second inequality applies $v_{s-1,i} \le (\mathcal{G}_T(s))^2$ in Lemma B.8 and the last inequality comes from $\sqrt{1-\beta_2}\mathcal{G}_T(s) \le \tilde{v}_{s,i}$. □

Proof of Lemma B.10. Applying the basic inequality and (35), for all $t \in [T], i \in [d]$,

$$\sum_{s=1}^t g_{s,i}^2 \le \sum_{s=1}^t \|\boldsymbol{g}_s\|^2 \le 2\sum_{s=1}^t\left(\|\bar{\boldsymbol{g}}_s\|^2 + \|\boldsymbol{\xi}_s\|^2\right) \le 2\mathcal{M}_T^2\left(\sigma_0^2 t + \sigma_1^2\sum_{s=1}^t\|\bar{\boldsymbol{g}}_s\|^p + \sum_{s=1}^t\|\bar{\boldsymbol{g}}_s\|^2\right). \tag{100}$$

Combining with Lemma B.2, we have

$$\begin{aligned}&\sum_{s=1}^t\|\bar{\boldsymbol{g}}_s\|^p \le \sum_{s=1}^t\left(\|\bar{\boldsymbol{g}}_1\| + \frac{LC_0\sqrt{d}s}{\sqrt{1-\beta_1/\beta_2}}\right)^p \le t\cdot\left(\|\bar{\boldsymbol{g}}_1\| + \frac{LC_0\sqrt{d}t}{\sqrt{1-\beta_1/\beta_2}}\right)^p \\ &\sum_{s=1}^t\|\bar{\boldsymbol{g}}_s\|^2 \le t\cdot\left(\|\bar{\boldsymbol{g}}_1\| + \frac{LC_0\sqrt{d}t}{\sqrt{1-\beta_1/\beta_2}}\right)^2.\end{aligned}$$

Further applying the definition of $\mathcal{F}_i(t)$ in Lemma B.3, it leads to $\mathcal{F}_i(t) \le \mathcal{F}(t), \forall i \in [d]$. Finally, since $\mathcal{F}(t)$ is increasing with t, we obtain the desired result. □

Proof of Lemma B.11. First, we have the following decomposition,

$$\mathbf{A.1} = -\sum_{s=1}^{t} \eta_s \left\| \frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}} \right\|^2 \underbrace{- \sum_{s=1}^{t} \eta_s \left\langle \bar{\boldsymbol{g}}_s, \frac{\boldsymbol{\xi}_s}{\boldsymbol{a}_s} \right\rangle}_{\mathbf{A}.1.1} + \underbrace{\sum_{s=1}^{t} \eta_s \left\langle \bar{\boldsymbol{g}}_s, \left(\frac{1}{\boldsymbol{a}_s} - \frac{1}{\boldsymbol{b}_s} \right) \boldsymbol{g}_s \right\rangle}_{\mathbf{A.1.2}}. \tag{101}$$

Since (35) holds, we could apply Cauchy-Schwarz inequality, Lemma B.9, and $\mathcal{G}_T(s) \le \mathcal{G}_T(t), \forall s \le t$ from (19) to obtain that for all $t \in [T]$,

$$\begin{aligned}
\mathbf{A.1.2} &\le \sum_{i=1}^{d} \sum_{s=1}^{t} \eta_s \left| \frac{1}{a_{s,i}} - \frac{1}{b_{s,i}} \right| \cdot |\bar{g}_{s,i} g_{s,i}| \le \sum_{i=1}^{d} \sum_{s=1}^{t} \eta_s \cdot \frac{\mathcal{G}_T(s)\sqrt{1-\beta_2}}{a_{s,i} b_{s,i}} \cdot |\bar{g}_{s,i} g_{s,i}| \\
&\le \frac{1}{4} \sum_{i=1}^{d} \sum_{s=1}^{t} \frac{\eta_s \bar{g}_{s,i}^2}{a_{s,i}} + (1-\beta_2) \sum_{i=1}^{d} \sum_{s=1}^{t} \frac{(\mathcal{G}_T(s))^2}{a_{s,i}} \cdot \frac{\eta_s g_{s,i}^2}{b_{s,i}^2} \\
&\overset{(30),(97)}{\le} \frac{1}{4} \sum_{s=1}^{t} \eta_s \left\| \frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}} \right\|^2 + \frac{\eta \mathcal{G}_T(t)\sqrt{1-\beta_2}}{1-\beta_1} \sum_{s=1}^{t} \left\| \frac{\boldsymbol{g}_s}{\boldsymbol{b}_s} \right\|^2.
\end{aligned}$$

Finally, combining with (37) for estimating **A.1.1**, we deduce the desired result in (39). □

Proof of Lemma B.12. Let us denote $\Sigma := \frac{\beta_1}{1-\beta_1} \langle \Delta_s \odot (\boldsymbol{x}_s - \boldsymbol{x}_{s-1}), \bar{\boldsymbol{g}}_s \rangle$ where Δ_s is defined in (20). We have

$$\begin{aligned}
\Sigma \le& \frac{\beta_1}{1-\beta_1} \cdot \left| \left\langle \Delta_s \odot \frac{\eta_{s-1} \boldsymbol{m}_{s-1}}{\boldsymbol{b}_{s-1}}, \bar{\boldsymbol{g}}_s \right\rangle \right| = \frac{\beta_1}{1-\beta_1} \cdot \left| \left\langle \left(\frac{\eta_s}{\boldsymbol{b}_s} - \frac{\eta_{s-1}}{\boldsymbol{b}_{s-1}} \right) \odot \boldsymbol{m}_{s-1}, \bar{\boldsymbol{g}}_s \right\rangle \right| \\
\le& \underbrace{\frac{\beta_1}{1-\beta_1} \cdot \left| \left\langle \left(\frac{\eta_s}{\boldsymbol{b}_s} - \frac{\eta_s}{\boldsymbol{a}_s} \right) \odot \boldsymbol{m}_{s-1}, \bar{\boldsymbol{g}}_s \right\rangle \right|}_{\Sigma_1} + \underbrace{\frac{\beta_1}{1-\beta_1} \cdot \left| \left\langle \left(\frac{\eta_s}{\boldsymbol{a}_s} - \frac{\eta_s}{\boldsymbol{b}_{s-1}} \right) \odot \boldsymbol{m}_{s-1}, \bar{\boldsymbol{g}}_s \right\rangle \right|}_{\Sigma_2} \\
&+ \underbrace{\frac{\beta_1}{1-\beta_1} \cdot \left| (\eta_{s-1} - \eta_s) \left\langle \frac{\boldsymbol{m}_{s-1}}{\boldsymbol{b}_{s-1}}, \bar{\boldsymbol{g}}_s \right\rangle \right|}_{\Sigma_3}.
\end{aligned} \tag{102}$$

Since (35) holds, we could apply Lemma B.9 and Young's inequality and then use (97), (30), $\beta_1 \in [0,1)$ and $\mathcal{G}_T(s) \le \mathcal{G}_T(t) \le \mathcal{G}_T(t) + \epsilon, \forall s \le t$,

$$\begin{aligned}
\Sigma_1 &\le \sum_{i=1}^{d} \frac{\beta_1}{1-\beta_1} \cdot \frac{\mathcal{G}_T(s)\eta_s\sqrt{1-\beta_2}}{a_{s,i} b_{s,i}} \cdot |\bar{g}_{s,i} m_{s-1,i}| \\
&\le \sum_{i=1}^{d} \frac{\eta_s}{8} \cdot \frac{\bar{g}_{s,i}^2}{a_{s,i}} + \frac{2\eta_s \beta_1^2 (1-\beta_2)}{(1-\beta_1)^2} \sum_{i=1}^{d} \frac{(\mathcal{G}_T(s))^2}{a_{s,i}} \cdot \frac{m_{s-1,i}^2}{b_{s,i}^2} \\
&\le \frac{\eta_s}{8} \left\| \frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}} \right\|^2 + \frac{2(\mathcal{G}_T(t)+\epsilon)\eta\sqrt{1-\beta_2}}{(1-\beta_1)^3} \left\| \frac{\boldsymbol{m}_{s-1}}{\boldsymbol{b}_s} \right\|^2.
\end{aligned} \tag{103}$$

Using the similar analysis for Σ_1, we also have

$$\begin{aligned}
\Sigma_2 &\le \sum_{i=1}^{d} \frac{\eta_s \beta_1}{1-\beta_1} \frac{\sqrt{1-\beta_2}}{a_{s,i} b_{s-1,i}} \cdot (\mathcal{G}_T(s) + \epsilon) \cdot |\bar{g}_{s,i} \cdot m_{s-1,i}| \\
&\le \frac{\eta_s}{8} \left\| \frac{\bar{\boldsymbol{g}}_s}{\sqrt{\boldsymbol{a}_s}} \right\|^2 + \frac{2\left(\mathcal{G}_T(t) + \epsilon\right)\eta\sqrt{1-\beta_2}}{(1-\beta_1)^3} \left\| \frac{\boldsymbol{m}_{s-1}}{\boldsymbol{b}_{s-1}} \right\|^2.
\end{aligned} \tag{104}$$

Then we move to bound the summation of Σ_3 over $s \in \{2, \cdots, t\}$ since $\boldsymbol{m}_0 = 0$. Recalling η_s in (30), we have the following decomposition,

$$\Sigma_3 \leq \underbrace{\frac{\eta\beta_1\sqrt{1-\beta_2^s}}{1-\beta_1}\left|\left(\frac{1}{1-\beta_1^{s-1}} - \frac{1}{1-\beta_1^s}\right)\left\langle \bar{\boldsymbol{g}}_s, \frac{\boldsymbol{m}_{s-1}}{\boldsymbol{b}_{s-1}}\right\rangle\right|}_{\Sigma_{3.1}}$$
$$+ \underbrace{\frac{\eta\beta_1}{(1-\beta_1)(1-\beta_1^{s-1})}\left|\left(\sqrt{1-\beta_2^{s-1}} - \sqrt{1-\beta_2^s}\right)\left\langle \bar{\boldsymbol{g}}_s, \frac{\boldsymbol{m}_{s-1}}{\boldsymbol{b}_{s-1}}\right\rangle\right|}_{\Sigma_{3.2}}. \tag{105}$$

Noting that $\|\bar{\boldsymbol{g}}_s\| \leq G_s \leq G_t, \forall s \leq t$. Then further applying Cauchy-Schwarz inequality and Lemma B.2,

$$\sqrt{1-\beta_2^s}\left|\left\langle \bar{\boldsymbol{g}}_s, \frac{\boldsymbol{m}_{s-1}}{\boldsymbol{b}_{s-1}}\right\rangle\right| \leq \sqrt{1-\beta_2^s}\|\bar{\boldsymbol{g}}_s\|\left\|\frac{\boldsymbol{m}_{s-1}}{\boldsymbol{b}_{s-1}}\right\| \leq \sqrt{d}G_t\sqrt{\frac{(1-\beta_1)(1-\beta_1^{s-1})}{(1-\beta_2)(1-\beta_1/\beta_2)}}.$$

Hence, summing $\Sigma_{3.1}$ up over $s \in [t]$, applying $\beta_1 \in (0,1)$ and noting that $\Sigma_{3.1}$ vanishes when $s = 1$,

$$\sum_{s=1}^t \Sigma_{3.1} \leq \frac{\sqrt{d}\eta G_t}{1-\beta_1} \cdot \sqrt{\frac{1-\beta_1}{(1-\beta_2)(1-\beta_1/\beta_2)}}\sum_{s=2}^t\left(\frac{1}{1-\beta_1^{s-1}} - \frac{1}{1-\beta_1^s}\right)$$
$$\leq \frac{\sqrt{d}\eta G_t}{\sqrt{(1-\beta_1)^3(1-\beta_2)(1-\beta_1/\beta_2)}}. \tag{106}$$

Similarly, using $\|\bar{\boldsymbol{g}}_s\| \leq G_s \leq G_t, \forall s \leq t$ and $1 - \beta_1^{s-1} \geq 1 - \beta_1$,

$$\frac{1}{1-\beta_1^{s-1}}\left|\left\langle \bar{\boldsymbol{g}}_s, \frac{\boldsymbol{m}_{s-1}}{\boldsymbol{b}_{s-1}}\right\rangle\right| \leq \frac{1}{1-\beta_1^{s-1}}\|\bar{\boldsymbol{g}}_s\|\left\|\frac{\boldsymbol{m}_{s-1}}{\boldsymbol{b}_{s-1}}\right\| \leq \sqrt{d}G_t\sqrt{\frac{1}{(1-\beta_2)(1-\beta_1/\beta_2)}}.$$

Hence, summing $\Sigma_{3.2}$ up over $s \in [t]$ and still applying $\beta_1 \in [0,1)$,

$$\sum_{s=1}^t \Sigma_{3.2} \leq \frac{\sqrt{d}\eta G_t}{1-\beta_1} \cdot \sqrt{\frac{1}{(1-\beta_2)(1-\beta_1/\beta_2)}}\sum_{s=2}^t\left(\sqrt{1-\beta_2^s} - \sqrt{1-\beta_2^{s-1}}\right)$$
$$\leq \frac{\sqrt{d}\eta G_t}{(1-\beta_1)\sqrt{(1-\beta_2)(1-\beta_1/\beta_2)}} \leq \frac{\sqrt{d}\eta G_t}{\sqrt{(1-\beta_1)^3(1-\beta_2)(1-\beta_1/\beta_2)}}. \tag{107}$$

Combining with (105), (106) and (107), we obtain an upper bound for $\sum_{s=1}^t \Sigma_3$. Summing (102), (103) and (104) up over $s \in [t]$, and combining with the estimation for $\sum_{s=1}^t \Sigma_3$, we obtain the desired inequality in (40). □

E Omitted proof in Appendix C

Proof of Lemma C.1. Recalling in (95), we have already shown that

$$\|\boldsymbol{x}_{s+1} - \boldsymbol{x}_s\| \leq \sqrt{d}\|\boldsymbol{x}_{s+1} - \boldsymbol{x}_s\|_\infty \leq \eta\sqrt{\frac{d}{(1-\beta_2)(1-\beta_1/\beta_2)}}, \quad \forall s \geq 1. \tag{108}$$

Applying the definition of $\boldsymbol{y}_s$ in (17), an intermediate result in (108) and $\beta_1 \in [0,1)$,[6]

$$\|\boldsymbol{y}_s - \boldsymbol{x}_s\| = \frac{\beta_1}{1-\beta_1}\|\boldsymbol{x}_s - \boldsymbol{x}_{s-1}\| \leq \frac{\eta}{1-\beta_1}\sqrt{\frac{d}{(1-\beta_2)(1-\beta_1/\beta_2)}}, \quad \forall s \geq 1. \tag{109}$$

[6] The inequality still holds for $s = 1$ since $\boldsymbol{x}_1 = \boldsymbol{y}_1$.

Recalling the iteration of $\boldsymbol{y}_s$ in (18) and then using Young's inequality

$$\|\boldsymbol{y}_{s+1}-\boldsymbol{y}_s\|^2 \le \underbrace{2\eta_s^2\left\|\frac{\boldsymbol{g}_s}{\boldsymbol{b}_s}\right\|^2}_{(*)} + \underbrace{\frac{2\beta_1^2}{(1-\beta_1)^2}\left\|\frac{\eta_s\boldsymbol{b}_{s-1}}{\eta_{s-1}\boldsymbol{b}_s}-\mathbf{1}\right\|_\infty^2\|\boldsymbol{x}_s-\boldsymbol{x}_{s-1}\|^2}_{(**)}.$$

Noting that $g_{s,i}/b_{s,i} \le 1/\sqrt{1-\beta_2}$ from (16), we then combine with (30) to have

$$(*) \le 2\eta_s^2 \cdot \frac{d}{1-\beta_2} \le \frac{2\eta^2 d}{(1-\beta_1)^2(1-\beta_2)}.$$

Applying Lemma B.1 where $\Sigma_{\max}^2 \le 1/\beta_2$ and (108),

$$(**) \le \frac{2\eta^2\beta_1^2\Sigma_{\max}^2 d}{(1-\beta_1)^2(1-\beta_2)(1-\beta_1/\beta_2)} \le \frac{2\eta^2 d}{\beta_2(1-\beta_1)^2(1-\beta_2)(1-\beta_1/\beta_2)}.$$

Summing up two estimations and using $0 \le \beta_1 < \beta_2 < 1$, we finally have

$$\|\boldsymbol{y}_{s+1}-\boldsymbol{y}_s\| \le \eta\sqrt{\frac{4d}{\beta_2(1-\beta_1)^2(1-\beta_2)(1-\beta_1/\beta_2)}}. \tag{110}$$

Combining with (108), (109) and (110), and using $0 \le \beta_1 < \beta_2 < 1$, we then deduce a uniform bound for all the three gaps. □

Proof of Lemma C.2. Under the same conditions in Lemma C.1, we have

$$\|\boldsymbol{y}_s-\boldsymbol{x}_s\| \le \frac{1}{L_q}, \quad \|\boldsymbol{y}_{s+1}-\boldsymbol{y}_s\| \le \frac{1}{L_q}.$$

Then, using the generalized smoothness in (8),

$$\begin{aligned}\|\nabla f(\boldsymbol{y}_s)\| &\le \|\nabla f(\boldsymbol{x}_s)\| + \|\nabla f(\boldsymbol{y}_s)-\nabla f(\boldsymbol{x}_s)\| \\ &\le \|\nabla f(\boldsymbol{x}_s)\| + (L_0 + L_q\|\nabla f(\boldsymbol{x}_s)\|^q)\|\boldsymbol{y}_s-\boldsymbol{x}_s\| \\ &\le \|\nabla f(\boldsymbol{x}_s)\| + \|\nabla f(\boldsymbol{x}_s)\|^q + L_0/L_q.\end{aligned}$$

We could use a similar argument to deduce the bound for $\|\nabla f(\boldsymbol{x}_s)\|$. Further, combining with $\mathcal{L}_s^{(x)}$ and $\mathcal{L}_s^{(y)}$ in (64), we could bound the generalized smooth parameters as

$$\begin{aligned}&L_0 + L_q\|\nabla f(\boldsymbol{x}_s)\|^q \le L_0 + L_q G_s^q = \mathcal{L}_s^{(x)}, \\ &L_0 + L_q\|\nabla f(\boldsymbol{y}_s)\|^q \le L_0 + L_q(\|\nabla f(\boldsymbol{x}_s)\| + \|\nabla f(\boldsymbol{x}_s)\|^q + L_0/L_q)^q = \mathcal{L}_s^{(y)}.\end{aligned} \tag{111}$$

We could then deduce the first two inequalities in (67). Finally, (68) could be deduced by using the same argument in the proof of [49, Lemma A.3]. □

Proof of Lemma C.3. Given any $\boldsymbol{x} \in \mathbb{R}^d$, we let

$$\tau = \frac{1}{L_0 + L_q\max\{\|\nabla f(\boldsymbol{x})\|^q, \|\nabla f(\boldsymbol{x})\|\}}, \quad \hat{\boldsymbol{x}} = \boldsymbol{x} - \tau\nabla f(\boldsymbol{x}).$$

From the definition of τ, we could easily verify that $\|\hat{\boldsymbol{x}}-\boldsymbol{x}\| = \tau\|\nabla f(\boldsymbol{x})\| \le 1/L_q$. Since f is (L_0, L_q)-smooth, we could thereby use the descent lemma in [49, Lemma A.3] such that

$$\begin{aligned}f(\hat{\boldsymbol{x}}) &\le f(\boldsymbol{x}) + \langle\nabla f(\boldsymbol{x}), \hat{\boldsymbol{x}}-\boldsymbol{x}\rangle + \frac{L_0 + L_q\|\nabla f(\boldsymbol{x})\|^q}{2}\|\hat{\boldsymbol{x}}-\boldsymbol{x}\|^2 \\ &= f(\boldsymbol{x}) - \tau\|\nabla f(\boldsymbol{x})\|^2 + \frac{(L_0 + L_q\|\nabla f(\boldsymbol{x})\|^q)\tau^2}{2}\|\nabla f(\boldsymbol{x})\|^2 \le f(\boldsymbol{x}) - \frac{\tau}{2}\|\nabla f(\boldsymbol{x})\|^2.\end{aligned}$$

Since $f(\hat{\boldsymbol{x}}) \ge f^*$, when $\|\nabla f(\boldsymbol{x})\| = 0$, the desired result is trivial. Let us suppose $\|\nabla f(\boldsymbol{x})\| > 0$.

Case 1 $\|\nabla f(\boldsymbol{x})\|^q > \|\nabla f(\boldsymbol{x})\|$

$$\frac{\tau}{2}\|\nabla f(\boldsymbol{x})\|^2 = \frac{\|\nabla f(\boldsymbol{x})\|^{2-q}}{2L_0/\|\nabla f(\boldsymbol{x})\|^q + 2L_q} \le f(\boldsymbol{x}) - f(\hat{\boldsymbol{x}}) \le f(\boldsymbol{x}) - f^*.$$

When $\|\nabla f(\boldsymbol{x})\|^q < L_0/L_q$, it leads to

$$\frac{\|\nabla f(\boldsymbol{x})\|^2}{4L_0} = \frac{\|\nabla f(\boldsymbol{x})\|^{2-q}}{4L_0/\|\nabla f(\boldsymbol{x})\|^q} \le \frac{\|\nabla f(\boldsymbol{x})\|^{2-q}}{2L_0/\|\nabla f(\boldsymbol{x})\|^q + 2L_q} \le f(\boldsymbol{x}) - f^*.$$

When $\|\nabla f(\boldsymbol{x})\|^q \ge L_0/L_q$, it leads to

$$\frac{\|\nabla f(\boldsymbol{x})\|^{2-q}}{4L_q} \le \frac{\|\nabla f(\boldsymbol{x})\|^{2-q}}{2L_0/\|\nabla f(\boldsymbol{x})\|^q + 2L_q} \le f(\boldsymbol{x}) - f^*.$$

We then deduce that

$$\|\nabla f(\boldsymbol{x})\| \le \max\left\{[4L_q(f(\boldsymbol{x}) - f^*)]^{\frac{1}{2-q}}, \sqrt{4L_0(f(\boldsymbol{x}) - f^*)}\right\}. \tag{112}$$

Case 2 $\|\nabla f(\boldsymbol{x})\|^q \le \|\nabla f(\boldsymbol{x})\|$ We could rely on the similar analysis to obtain that[7]

$$\|\nabla f(\boldsymbol{x})\| \le \max\left\{4L_q(f(\boldsymbol{x}) - f^*), \sqrt{4L_0(f(\boldsymbol{x}) - f^*)}\right\}. \tag{113}$$

Combining (112) and (113), we then deduce the desired result. □

Proof of Lemma C.5. Recalling (95), we then obtained that when $\eta = \tilde{C}_0\sqrt{1-\beta_2}$,

$$\|\boldsymbol{x}_s - \boldsymbol{x}_{s-1}\| \le \sqrt{d}\|\boldsymbol{x}_s - \boldsymbol{x}_{s-1}\|_\infty \le \tilde{C}_0\sqrt{\frac{d}{1-\beta_1/\beta_2}}. \tag{114}$$

Noting that when (66) holds, we have

$$\begin{aligned}\|\bar{\boldsymbol{g}}_s\| &\le \|\bar{\boldsymbol{g}}_{s-1}\| + \|\bar{\boldsymbol{g}}_s - \bar{\boldsymbol{g}}_{s-1}\| \le \|\bar{\boldsymbol{g}}_{s-1}\| + (L_0 + L_q\|\bar{\boldsymbol{g}}_{s-1}\|^q)\|\boldsymbol{x}_s - \boldsymbol{x}_{s-1}\| \\ &\le \|\bar{\boldsymbol{g}}_{s-1}\| + \tilde{C}_0\mathcal{L}_{s-1}^{(x)}\sqrt{\frac{d}{1-\beta_1/\beta_2}} \le \|\bar{\boldsymbol{g}}_1\| + \tilde{C}_0\sqrt{\frac{d}{1-\beta_1/\beta_2}}\sum_{j=1}^{s-1}\mathcal{L}_j^{(x)}.\end{aligned}$$

Using $\mathcal{L}_j^{(x)} \le \mathcal{L}_t^{(x)}, \forall j \le t$, we have

$$\sum_{s=1}^t \|\bar{\boldsymbol{g}}_s\|^p \le \sum_{s=1}^t\left(\|\bar{\boldsymbol{g}}_1\| + \tilde{C}_0\sqrt{\frac{d}{1-\beta_1/\beta_2}}(s-1)\mathcal{L}_t^{(x)}\right)^p \le t\left(\|\bar{\boldsymbol{g}}_1\| + t\tilde{C}_0\mathcal{L}_t^{(x)}\sqrt{\frac{d}{1-\beta_1/\beta_2}}\right)^p.$$

Similarly, we also have

$$\sum_{s=1}^t \|\bar{\boldsymbol{g}}_s\|^2 \le t\left(\|\bar{\boldsymbol{g}}_1\| + t\tilde{C}_0\mathcal{L}_t^{(x)}\sqrt{\frac{d}{1-\beta_1/\beta_2}}\right)^2.$$

Further combining with $\mathcal{F}_i(t)$ in Lemma B.3 and $\mathcal{J}(t)$ in (71),

$$\mathcal{F}_i(t) \le 1 + \frac{1}{\epsilon^2}\sum_{s=1}^t \|\boldsymbol{g}_s\|^2 \le \mathcal{J}(t), \quad \forall t \in [T], i \in [d].$$

□

[7] We refer readers to see [51, Lemma A.5] for a detailed proof under this case.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading “NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: [NA]

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: [NA]

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [NA]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.